THE UNIVERSITY
of EDINBURGH

Improving the Predictive Performance of Longitudinal Risk Models for UK SMEs

Haoye Liang

THE UNIVERSITY
of EDINBURGH

Degree of Doctor of Philosophy

in Management Science

The University of Edinburgh

2019

Declaration

In accordance with the University of Edinburgh Regulations for Research Degrees, the author declares that:

1. This thesis has been composed by the author.
2. It is the result of the author's own original research.
3. It has not previously been submitted for any other degree or professional qualification.
4. Preliminary results of this research were presented at international conferences and workshops as per attached list of Refereed Conference Papers.

Signed:

Date:

Acknowledgements

My deepest gratitude goes to my supervisors, Professor Jake Ansell. He pointed out the research direction of SMEs and gave me profound suggestions. His enthusiasm, patience and expertise have positively added to my PhD experience.

I would like to thank my loving family, my mother, father and brother, for all the support they provided me throughout my PhD period. I would not have finished this thesis without their enormous encouragement and unconditional love.

Abstract

In 2008, the whole world was a picture of economic depression. During the credit crisis, the viability of Small and Medium-sized Enterprises (SMEs) has been profoundly jeopardised because of their vulnerability. After the credit crisis, raising the awareness of risk management in the banking sector has been needed. The launch of the Basel III regulations proposes a more stringent requirement on capital and liquidity to promote stability in the financial system. Regarding credit risk, the probability of defaults (PDs) models essentially remains unchanged. On the other hand, the expected credit loss (ECL) model under International Financial Reporting Standard (IFRS) 9 proposed by International Accounting Standard Board (IASB) is expected to bring significant influence on SMEs and Banking sectors since there is an increase of loan loss provision undoubtedly.

This thesis aims to explore the performance of SMEs due to the fundamental role played in a country's economic development. A large dataset used in this thesis includes 79 characteristics of UK SMEs from 2007 to 2010. The SMEs were pigeonholed into start-ups (growing businesses) and non-start-ups (developed businesses) considering their different behaviour during the credit crisis. However, the dataset contains a substantial number of incomplete observations, and the analysis of such dataset is a handicap. In light of this, Multiple Imputation by Chain Equations (MICE), a state-of-the-art and flexible technique, has been employed to deal with missing data. Although this technique is widely used in the medical field but not in credit risk modelling, it takes into account the uncertainty within the process of combing multiple imputed dataset to produce estimated coefficient, and each type of variables has its specific model for imputation.

Once getting over the missing data problem, cross-section analysis is followed to build up the credit risk model. Logistic regression and shrinkage regression are used to analyse the relationship among the selected variables and Generalised Additive Models (GAM) is performed to capture their non-linear relationship, derive a direct marginal trend and plot how explanatory variables influence SMEs

performance. Subsequently, time effects are accounted for by employing a panel model controlling the time effect using year dummy variables or macroeconomic variables. It can be found that the panel data models with firm-specific and macroeconomic variables are preferred as the AUROC is at least as other models, especially during the credit crisis.

Again, the ex-post regulations, the 12-month ECL may not capture a significant increase in credit risk if the economic downturn is expected to occur at a later stage. The lifetime ECL captures this downturn and will, therefore, identify a significant increase in credit risk sooner. The panel data models are believed to capture the change in the macroeconomics during the credit risk and are appropriate to apply to meet Basel III and IFRS 9 requirements as these regulations require to consider the economic cycle.

CONTENTS

LIST OF FIGURES ..... **I**

LIST OF TABLES ..... **III**

LIST OF ABBREVIATIONS ..... **V**

CHAPTER 1 INTRODUCTION **1**

1.1 BACKGROUND 1

1.1.1 SME's Definition *1*

1.1.2 Importance of SMEs *2*

1.1.3 The Impact from the Credit Crisis of 2008-09 in the UK *2*

1.1.4 Causes and Remedial Actions *5*

1.2 MOTIVATION FOR THE RESEARCH 7

1.3 RESEARCH OBJECTIVES AND QUESTIONS 8

1.4 CONTRIBUTIONS 9

1.5 STRUCTURE OF THE THESIS 11

CHAPTER 2 LITERATURE REVIEW **13**

2.1 INTRODUCTION 13

2.2 MISSING DATA 14

2.2.1 Missing Data Mechanisms *15*

2.2.2 Advanced Methods for Missing Data *19*

2.3 CREDIT RISK MODELS 26

2.3.1 Definition *26*

2.3.2 Business Failure Prediction *27*

2.3.3 Credit Risk Models for SMEs *31*

2.4 FRAMEWORK OF BASEL ACCORD 34

2.4.1 Basel I Accord *36*

2.4.2 Basel II Accord *38*

2.4.3 Basel III Accord *44*

2.4.4 Basel Accord about SMEs *47*

2.5 INTERNATIONAL FINANCIAL REPORTING STANDARD (IFRS) 9 51

2.5.1 From Incurred Loss (IAS 39) to Expected Credit Loss (IFRS 9) *52*

2.5.2 Impairment Model Under IFRS 9 56
2.5.3 Significant Increase in Credit Risk(SICR) 59
2.5.4 12-month ECL and Lifetime ECL 61
2.5.5 The Impact of IFRS 9 on Banks and SMEs 64
2.6 Basel Accord and IFRS 9 66
2.6.1 Rating Philosophies 66
2.6.2 Capital Ratio and Provisions 70
2.7 Definition of Default 70
2.8 Summary 72
CHAPTER 3 METHODOLOGY 75
3.1 Introduction 75
3.2 Multiple Imputation by Chain Equations (MICE) 77
3.2.1 MICE steps 78
3.2.2 Combining Rules 80
3.2.3 Imputation Diagnostic Measure 82
3.2.4 Number of Imputations 84
3.2.5 Non-normally Distributed Variables 87
3.2.6 Auxiliary Variable 88
3.3 Independent Variables Selection 89
3.3.1 Weight of Evidence and Information Value 89
3.3.2 Analysis with Multiple Datasets (after Imputation) 91
3.3.3 Empirical 93
3.4 Logistic Regression 95
3.5 Shrinkage Regression 97
3.6 Generalized Additive Models (GAMs) 101
3.7 Panel Data Analysis 106
3.7.1 Fixed versus Random Effects 108
3.7.2 Macroeconomic Variables (MVs) 110
3.8 Model Performance 111
3.9 Summary 112
CHAPTER 4 DATA DESCRIPTION 115
4.1 Introduction 115

4.2 SAMPLE DATA 115
4.3 "BAD" RATES 116
4.4 VARIABLES EXPLANATION 118
4.4.1 1992 SIC Code 118
4.4.2 Regions 122
4.4.3 Time since Last Derogatory Data Item (Months) 125
4.4.4 Proportion of Current Directors to Previous Directors in the Last Year 128
4.4.5 Time since Last Annual Return 128
4.5 SUMMARY 131

CHAPTER 5 RESULTS 133

5.1 IMPUTATION 133
5.1.1 Results of MICE Imputation 133
5.1.2 Imputations of empirical variables 143
5.1.3 Variable Confirmation 151
5.1.4 Summary 152
5.2 CROSS-SECTION MODELS 153
5.2.1 Logistic Regression with Weight on the Stacked Dataset 153
5.2.2 Logistic Regression with WoE Data 156
5.2.3 Shrinkage Regression with WoE Data 158
5.2.4 Generalised Additive Model (GAM) with Imputed Dataset 160
5.2.5 Model Performance of Cross-section Analysis 179
5.3 PANEL MODELS 181
5.3.1 Macroeconomic Variables (MVs) 181
5.3.2 Explanatory Models 186
5.4 SUMMARY 190

CHAPTER 6 CONCLUSIONS AND DISCUSSIONS 192

6.1 RESEARCH QUESTIONS ANSWERED 193
6.2 LIMITATION AND SUGGESTION FOR FURTHER RESEARCH 197
6.2.1 Dealing with Missing Data 198
6.2.2 Data-quality Reject Inference 199
6.2.3 Credit Risk Modelling 200

6.2.4 Micro-Enterprises.....200

LIST OF FIGURES

Figure 1-1 The United Kingdom GDP growth rate 3
Figure 1-2 The vicious circle for SMEs lending..... 4
Figure 1-3 The number of bankruptcies in the United Kingdom..... 5
Figure 2-1 Overview of the MI procedure..... 23
Figure 2-2 Capital requirement for IRB approach under Basel II taken from 41
Figure 2-3 Loan loss recognition under alternative accounting regimes taken from G. u. Gebhardt and Novotny-Farkas (2011)..... 54
Figure 2-4 Development of provisions under IFRS 9 and IAS 39 taken from (Frykström and Jieying, 2018) 55
Figure 2-5 PD of TTC versus PIT over the business cycle taken from 68
Figure 3-1 Development flow from raw data to model building. 75
Figure 3-2 Figure that help explains why lasso can select predictors, taken from Jerome Friedman, et al. (2001)..... 99
Figure 3-3 ROC of estimated default probability 112
Figure 4-1 Frequency percentage plots of time since last derogatory data item (months).....126
Figure 4-2 Frequency percentage plots of time since last derogatory data item (months) after removing the missing group.....127
Figure 4-3 Frequency percentage plots of Proportion of current directors to previous directors in the last year 129
Figure 4-4 Frequency percentage plots of time since last annual return 130
Figure 5-1 Plots of convergence and distribution comparison 140
Figure 5-2 Bar chart of observed and imputed values of last derogatory item ..141
Figure 5-3 Bar chart of observed and imputed values of 1992 SIC code.....143
Figure 5-4 Convergence plots of variables over 50% missing rate for start-ups 147
Figure 5-5 Convergence plots of variables over 50% missing rate for non-start-ups 148
Figure 5-6 Density plot of continuous variables 150

Figure 5-7 GAM - Proportion of current directors to previous directors in the last year.....162
Figure 5-8 GAM - Oldest age of current directors/proprietors supplied (years).164
Figure 5-9 GAM - Number of directors holding shares.....165
Figure 5-10 GAM - Total value of judgements in the last 12 months 166
Figure 5-11 GAM - Time since last derogatory data item (months) 168
Figure 5-12 GAM - Lateness of accounts 169
Figure 5-13 GAM - Time since last annual return 170
Figure 5-14 GAM - Total assets.....170
Figure 5-15 GAM - No. of ‘current’ directors 173
Figure 5-16 GAM - PP worst (company DBT - industry DBT) in the last 12 months 173
Figure 5-17 GAM - Total value of judgements in the last 12 months 174
Figure 5-18 GAM - Time since last derogatory data item (months) 175
Figure 5-19 GAM - Lateness of accounts 176
Figure 5-20 GAM - Time since last annual return 176
Figure 5-21 GAM - Total fixed assets as a percentage of total assets.....177

LIST OF TABLES

Table 2-1 Summary of credit risk models for SMEs 33
Table 2-2 History of Basel Accord 35
Table 2-3 Risk Weights for On-Balance Sheet Items 37
Table 2-4 Risk Weights as a Percent of Principal for Exposures to Countries, Banks, and Corporations Under Basel ll's Standardized Approach 40
Table 2-5 Relationship between PD and WCDR for firm, bank and retail exposure 43
Table 2-6 Basel III Accord minimum capital requirement 47
Table 2-7 Overview of the general IFRS 9 impairment approach 58
Table 3-1 Summary of general MICE imputation models 78
Table 3-2 Missing rates (%) in the dataset 94
Table 4-1 Frequency of UK SMEs data 117
Table 4-2 Percentage (%) table of industry sectors 120
Table 4-3 "bad" rate (%) across different industries 121
Table 4-4 Percentage table of regions 123
Table 4-5 'bad' rate across different regions 124
Table 5-1 Pooled logistic regression results (the result of step 3 of MICE) 135
Table 5-2 Summary statistics of the observed and imputed data for the incomplete variables in the analysis model selected by Rubin's rules 139
Table 5-3 Pooled results of Start-ups 145
Table 5-4 Pooled results of Non-start-ups 146
Table 5-5 Independent variables used for prediction 151
Table 5-6 Coefficients of logistic regression using a stacked dataset of start-ups with weights 155
Table 5-7 Coefficients of logistic regression using a stacked dataset of non-start-ups with weights 156
Table 5-8 Coefficient estimates for logistic regression of start-ups data with woe transformation 157

Table 5-9 Coefficient estimates for logistic regression of non-start-ups data with woe transformation 158
Table 5-10 Coefficient estimates for start-ups for lambda.1se.....159
Table 5-11 Coefficient estimates for non-start-ups for lambda.1se.....159
Table 5-12 Effective degrees of freedom and approximate significance of each GAM smoothed term of start-ups.....162
Table 5-13 Effective degrees of freedom and approximate significance of each GAM smoothed term of non-start-ups.....172
Table 5-14 AUROC on the test sample.....180
Table 5-15 UK Macroeconomic data from 2007 to 2010.....182
Table 5-16 Correlation of UK Macroeconomic data 183
Table 5-17 Analysis of individual MVs 185
Table 5-18 Start-up random effect panel data models parameter estimation ...188
Table 5-19 Non-start-up random effect panel data models parameter estimation 189
Table 5-20 AUROC of panel models.....189

LIST OF ABBREVIATIONS

AUROC	Area Under Receiver Operation Characteristics
CPI	Consumer Price Index
ECL	Expected Credit Loss
FE	Fixed Effect
FVA	Fair Value Accounting
GAM	Generalized Additive Models
GDP	Gross Domestic Product
IASB	International Accounting Standard Board
IAS 39	International Accounting Standard 39
IFRS 9	International Financial Reporting Standard 9
IRB	Internal-Rating Approach
IV	Information Value
MAR	Missing at Random
MCAR	Missing Complete at Random
MI	Multiple Imputation
MICE	Multiple Imputation by Chain Equations
ML	Maximum Likelihood
MNAR	Missing Not at Random
PD	Probability of Default
PMM	Predictive Mean Matching
RE	Random Effect
SA	Standardised Approach
SMEs	Small and Medium-Sized Enterprises
WoE	Weight of Evidence
FMI	Fraction of Missing Information

CHAPTER 1 INTRODUCTION

1.1 Background

1.1.1 SME's Definition

Most small and medium-sized enterprises (SMEs) often require access to finance to be sustainable (Duan, Han, & Yang, 2009; T. Beck, Demirgüç-Kunt, & Pería, 2010; Irwin and Scott, 2010; Hussain, Salia, & Karim, 2018). To receive funding, the European Union requires an explicit definition of SMEs and classification of their creditworthiness. Currently, there is no consensus on the definition in global taking into account various quantitative and firm characteristics. Yet, the majority of them are non-subsidiary, and independent firms which employ less than a given number of employees, such as 500 in the United States in general.

In the UK, being an SME needs to meet two out of three conditions: turnover (less than £25m), employees (less than 250), and gross assets (less than £12.5m)[1]. Given that the source of the datasets in this research, the European Union definition will be used. In 1996, the European Union standardized the term "SME" and defined it arising from Basal Accord[2]. Thus, the quantitative upper limits of the SME have:

- Number of employees should be no more than 250
- Either total turnover is less than €50 million, or a balance sheet total is less than €43 million.

As soon as one of the criteria is exceeded, the European Union no longer classifies the company as an SME but instead moves it into the large enterprise classification.

[1] source: https://www.gov.uk/government/collections/mid-sized-businesses
[2] source: https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en

Specifically, a small-sized enterprise is defined as the number of employees must be between 10 and 50, and the annual turnover is between €2 million and €10 million. A medium-sized enterprise is above the condition, and micro-sized business is below the condition.

1.1.2 Importance of SMEs

SMEs are critical to the further boost of a nation's economic (Y. Ma and Lin, 2010; Venkatesh and Muthiah, 2012; Calabrese, Andreeva, & Ansell, 2017; Kumar et al., 2018). According to a report from the Department for Business Innovation and Skills indicated that in 2017 data, 99.9% of 5.7 million private sector business were SMEs. Total employment in SMEs was 16.1 million, which was 60% of all private-sector employment in the UK. The combined annual turnover of SMEs was £1.9 trillion, which was 51% of all private sector turnover in the UK (Department for Business, 2017). Pagano and Pica (2012) found that there is a significant and positive relationship between financial development and job creation in developed countries, which partly happens through expanding SMEs finance (T. Beck, 2013). Besides, SMEs encourage competition and bring new ideas that challenge the status quo, which stimulus, in turn, motivates others to adapt. Needless to say, they should be encouraged to blossom. These evidences have shown that SMEs are the backbone of the UK economy by contributing to employment opportunity, innovative development and economic growth.

1.1.3 The Impact from the Credit Crisis of 2008-09 in the UK

Around one decade ago, the highly risky subprime mortgage market provided the financial sector with the potential for growth, yet it became problematic as the number of foreclosures increased, leading to financial companies facing severe stress. In September 2008, the insolvency of US investment bank Lehman Brothers triggered one of the worst global economic crises since 1929. The financial crisis became the harshest problem in the world during this period and generated a sudden change in the economic policy of the UK. The Bank

of England gradually cut the interest rate from 5.5% in 2007 to 0.5% in 2009. 0.5% was the lowest since the central bank was established in 1694.

The worldwide financial turmoil that began in 2007 triggered the first run on a British bank since 1866 and a near meltdown in the banking system 12 months later. The credit crunch, the effects of which have been amplified by the bursting of the UK's decade-old house price bubble, has taken a severe toll on the economy (Hodson and Mabbett, 2009). Figure 1-1 displays that UK GDP showed a negative growth rate between 2008 and 2009, according to the World Bank national accounts data and OECD National Accounts data files since 2000[3]. In addition to the unemployment rate, it increased from 5.2% in 2007 to 7.9% in 2009[4].

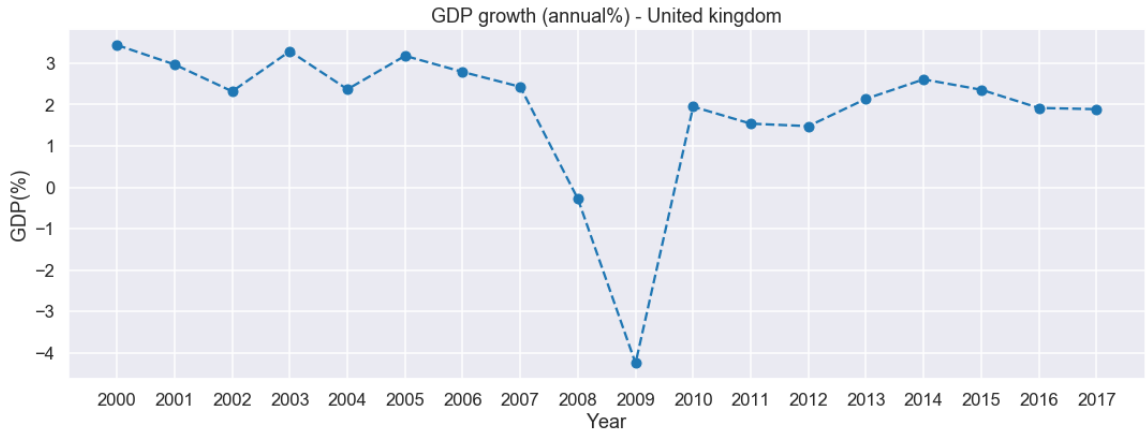


Figure 1-1 The United Kingdom GDP growth rate

After the outbreak of the credit crisis, banks have been urged to increase its equity to maintain the minimum requirement of capital ratios (Iqbal and Kume, 2014), and to increase lending to SMEs by the UK government. Loans to the private sector in the United Kingdom reached an all-time high of 2,813,033

[3] source: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?contextual=min&end=2018&locations=GB&start=2000&view=chart
[4] source: https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms

GBP million in the first quarter of 2009[5]. Nevertheless, the level of difficulty to obtain funding from the banking sector increased. A sharp fall of bank base rate seem to be ineffectual: inter-bank lending rates kept high and volumes low (Hodson and Mabbett, 2009). The roots of the financial crisis lay in overvalued assets (houses), mainly those backed by mortgages. As these assets began to lose value, it was unclear who owned them and so was exposed to the losses. Banks became more cautious (lack of confidence) when they issued loans to SMEs, or even unwilling to lend to other banks, and restrictions in lending spread through into a wider economy, thus occurring the 'credit crunch' (Fosberg, 2012; N. Lee, Sameen, & Cowling, 2015), or say 'liquidity crisis' resulted from a decrease of the supply of loans to either financial or non-financial firms.

During the period of the credit crisis, it is expected that some SMEs were very frail. Hence some SMEs were severely hit because they are vulnerable (Orton, Ansell, & Andreeva, 2017), and high credit risk (Jacobson, Lindé, & Roszbach, 2005) thus bring about a vicious spiral between the SMEs performance and credit amounts, as shown in Figure 1-2.

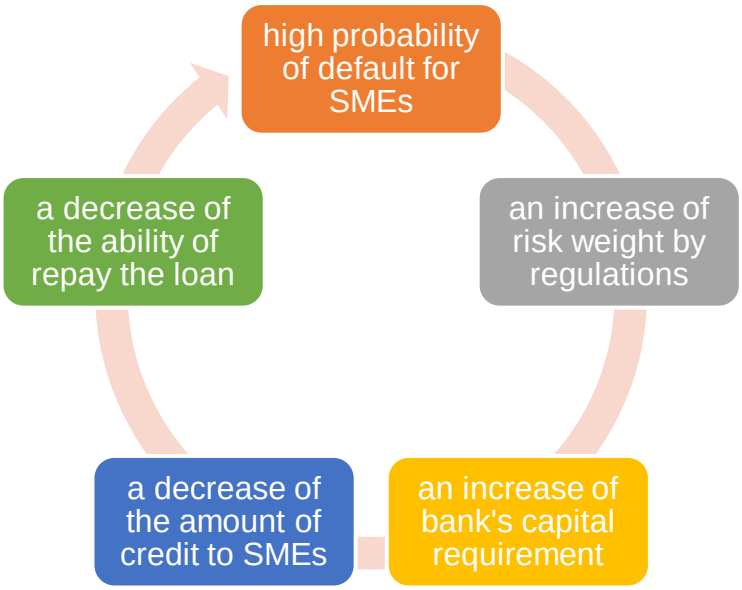


Figure 1-2 The vicious circle for SMEs lending

[5] source: https://tradingeconomics.com/united-kingdom/loans-to-private-sector

The amount of monthly lending to SMEs in the UK dropped from 991 million to 566 million between 2008 and 2010. Finally, this situation resulted in substantial growth in bankruptcies of SMEs due to a lack of funding. The UK SMEs default rate was 6.9% in 2007, followed by a dramatic increase of 11.8% in 2008, and reached the top at 16.1% in 2009, finally decreased to 11.9% in 2010. Figure 1-3 below provides the number of UK bankruptcies, and it is clear that the number rocketed to a peak in 2009[6].

Figure 1-3 The number of bankruptcies in the United Kingdom

1.1.4 Causes and Remedial Actions

During the crisis period, a number of SMEs already weakened by the collapse of economic growth and announced zero profits or even losses. An increasing number of bankruptcies were announced, and the financial distress spread across different types of firms and industries. The massive loss caused by the bankruptcies led to considerable criticism of the efficiency of financial institutions, which is partly because of the inappropriate evaluation of credit risk driving inaccurate credit ratings of both homeowners and bonds. This economic crisis aroused awareness of the importance and influence of credit

[6] source: https://tradingeconomics.com/united-kingdom/bankruptcies

risk management. In specific, banking systems should evaluate credit risk more precisely and effectively. Furthermore, the experiences learnt from the financial crisis required the redesign of the rules governing financial market participants and banking supervision.

Most governments were forced to implement rescue plans so that the government of the G20 states met in Pittsburgh, the USA in 2009, and decided to improve the resilience of the financial markets using new regulations on banking supervision. The Basel Committee on Banking Supervision was tasked with implementing this decision. On 16 December 2010, it published the new set of rules known as Basel III (BCBS, 2011). The global capital framework and new capital buffers require financial institutions to hold more capital and higher quality of capital than under previous Basel II rules. The Basel II and Basel III Accord both try to address internal risk management tools that focus on SMEs. The revised version of the Basel III regulations was issued by the Basel Committee on Banking Supervision providing Basel regulatory framework for capital adequacy of banks. Significant changes in comparison to Basel II are the enhancement of quality of regulatory capital as well as its size. The calculation methods of necessary minimum regulatory capital requirement for every exposure under the current regulations, however, are still based on the first pillar of the preceding Basel II. Under the current Basel regulations, there are greater incentives for banks to adopt the Internal Ratings-Based Approach (IRBA) to determine their regulatory capital requirement. Banks can use external scoring or rating assessments, which is known as an external rating approach. However, they apply to only several of the largest business entities (Nehrebecka and Polski, 2016), and the specific research of SMEs with regard to credit risk modelling is still ambiguous in the position of corporate or retail exposures although SMEs' research has had significant attention.

More recently, the international financial reporting standard IFRS 9 Financial instruments, which came into force on 1st of January 2018, emphasises and

deepens requirements in the area of credit risk analysis and management even more. In a sense, it also created a stronger link between credit risk and accounting, and significantly impact on the banks' financial results. IFRS 9 replaces the international accounting standard IAS 39 Financial Instruments: Recognition and Measurement mainly because the biggest weakness of IAS 39 has proved to be the mechanism of calculating impairment (credit losses) associated with financial assets and their loss allowances accounting. The deficient impairment framework of IAS 39 was the most persuasive reason for it to be replaced by IFRS 9 (Vaněk and Hampel, 2017).

1.2 Motivation for the Research

The status of SMEs in social development is irreplaceable, but they are opaquer, and they lack trustworthy external ratings compared with listed companies because their financial or operation situations are not exposed to the public. It was found that SMEs are less likely to be able to obtain bank loans than large firms especially during the last crisis period; instead, they rely on internal funds, or cash from friends and family, to launch and initially run their enterprises.

In addition, the existence of missing data is unavoidable in the social sciences, and this happens to the field of credit scoring as well. On the one hand, most researches are likely to delete those incomplete observations or apply simple single imputation, such as mean substitution, when modelling credit risk so that the accuracy of the model will be impaired. On the other hand, traditionally, the binning method, such as weight of evidence (WoE), is widely used and has satisfactory achievement in the field but it is difficult to interpret when dealing with non-linearity. Therefore, one might have to deal with this problem prior to build up credit risk model.

Furthermore, SMEs are treated differently from big companies when making a decision to grant credit. Verbano and Venturini (2013) pointed out that there

have been little researches to improve the survival possibility of SMEs and create values considering risk management. Basel II Accord proposed a special treatment to SMEs different risk weights are given according to different categories (corporate or retail exposure), and retail credit and loans to SMEs will receive a different treatment than corporate loans and will require less regulatory capital for given default probabilities. The main reason for this different treatment is that small business loans and retail credit are generally found to be less sensitive to systematic risk. Their risk of default is thought to be mainly of an idiosyncratic nature and, as a result, default probabilities are assumed to be more weakly correlated when compared with corporate loans (Jacobson, et al., 2005). During the credit crisis, Banks distrusted of SMEs' ability to repay so SMEs received only limited help, even if the state encouraged them to lend.

In summary, this research would like to improve the predictive performance of the credit risk model focusing on SMEs so that it is able to conducive to maintaining social stability and development.

1.3 Research Objectives and Questions

This research aims to improve the predictive performance regarding SMEs based on the data from the last credit crisis. The thesis concentrates on UK SMEs credit risk modelling using the data from 2007 to 2010 in relation to Basel Accord and IFRS 9. The research has the following objective:

1. To explore missing data approaches that are suitable for SMEs data considering that there is a large proportion of missing data and mixed type of variables
2. To develop more accurate credit risk models for SMEs that can improve the ‘bad’ rate prediction especially during the credit crisis period and evaluate these model’s predictive accuracy
3. To discuss the impact of IFRS 9 on the bank sectors and SMEs

These objectives will be investigated by a set of research questions:

- Except for WoE, is there another way to handle missing data better than the convention methods such as listwise deletion?
- Does logistic regression using an alternative method provide an acceptable prediction accuracy for SMEs probability of default? In addition to logistic regression, is there a better approach to improve the prediction accuracy?
- Is there any non-linear relation between independent variables and SMEs' performance? How to model with the non-linear effects?
- During the credit crisis, there was a significant shock on macroeconomic, do these effects have a marked impact on the viability of SMEs? How to model the SME' performance due to the change in the macroeconomic environment?
- How the implementation of IFRS 9 affects banks and SMEs?

1.4 Contributions

To investigate the SMEs performance requires to cope with missing values problems. This research provides an insight into multiple imputation by chain equations (MICE), and its ability to deal with a large proportion of missing data with and mixed type of variables in the SMEs dataset in order to try to retain the statistical power and recover the useful information. Since the binning method is widely used in credit scoring, the comparison of the performance to handle missing data will be of value to researchers and practitioners.

This research explores the viability of SMEs during the last credit crisis. The implementation of Basel Accord II has inspired a number of studies of default probabilities by banks to lower the capital. Industry classifier (logistic regression) is built as a benchmark to make comparisons to other methods used in this research. In addition, the marginal effects of individual independent variables and non-linear effect are also examined through the generalized additive model which is not common in credit scoring in practice.

This task becomes a particularly challenging and vital issue for banks and financial institutions to access the performance of SMEs. Therefore, financial credit risk assessments have become a measure to assess SMEs credit access or potential business failures, so banks and financial institutions can act as early as possible before the actual financial crisis. Finally, by considering the time effect with panel data models, it is also used to observe the SMEs performance from the multi-period aspect instead of a single period by capturing the change in economic conditions. The performance of panel models is at least as the cross-section model with firm-specific variables only especially for non-start-up during the credit crisis. Panel models are ideal for modelling probability of default as Basel Accord and IFRS 9 require considering the macroeconomic factors.

Predicting the potential loss is the root issue when modelling credit risk. So, the credit quality of a borrower does not only depend on the default probability, but also on the exposure at default and the loss given default. However, most studies concentrated on the prediction of the default probability historically. Besides, Basel II differentiates between the Foundation and the Advanced Internal-Rating Approach, where for the Foundation Approach banks only have to estimate default probabilities. Due to these reasons and data unavailability for the exposure at default and the loss given default, the majority study will focus on default probabilities prediction. Researchers are more interested in credit risk modelling while practitioners focus on the details of implementation, whether the model in this research is better, whether the model is too complex to be used in practice, and any profits/losses brought from changing the models.

In summary, this research delves into missing data handling and different methodologies to model SMEs credit risk, especially with GAMs to dealing with non-linearity in order to improve the predictive performance of SMEs.

1.5 Structure of the Thesis

The following provides a brief illustration of the rest of this thesis:

Chapter two Literature Review:
Missing data is unavoidable, and it has a material effect on the accuracy of the statistical analysis. Approaches dealing with missing data are reviewed, and multiple imputation by chain equation (MICE) is recommended to deal with missing data in credit scoring areas due to its flexibility. Credit risk modelling is also reviewed. Basel II Accord changed the way banks calculated their capital requirement to address their credit risk by using the standardised approach (SA) or internal-rating approach (IRB), which could be possible to lower the capital requirement. Although Basel III came into force in 2009, most of the credit-risk-related regulations are the same as Basel II. Incurred losses model from IAS 39 has been heavily criticised since the outbreak of the last credit crisis due to its inefficiency and has been replaced by a forward-looking standard: IFRS 9. There are differences between the Basel Accord and IFRS 9, yet they both provide a guideline to reduce risk. Additionally, the Basel Accord and IFRS 9 have a special treatment for SMEs.

Chapter three Methodology:
The main feature of this chapter is to look at the performance of classification methodology used in this thesis, and there are two treatments of the data: binning transformation and stacked imputed data. The combination of multiple imputed dataset to estimate coefficient may incorrectly specified the distribution and lead to the overfitting problem. An alternative to estimating coefficient is to stack data with weight instead of multiple imputed dataset. Logistic regression, shrinkage regression, generalized additive model (GAM) and panel model are introduced to build credit risk models. The predictive power of models is validated by using Receiver Operation Characteristics (ROC) plots and Area under ROC (AUROC).

Chapter four Data Description:

This chapter describes the employed dataset in detail. The dataset records mainly four types of information including general information, directors' information, and previous relevant credit history and accounting information. The SMEs' creditworthiness is labelled as good or bad. It is necessary to sample the dataset considering the time spent in MICE process and remove the observations which do not meet the SMEs definitions. Besides, splitting SMEs as start-ups and non-start-ups SMEs lays down the foundation for further credit risk modelling because of an initially marked difference on 'bad' rate for the two segments. Next, to analyse industry performance, and regional performance and other factors for start-ups and non-start-ups can initially reveal the reasons why there is a significant difference in the default rate. A brief summary will be provided in the last section.

Chapter five Results:

In this chapter, the imputation and models analysis result will be provided. The first section shows the analysis consequence of imputation by providing the corresponding parameters and plotting and comparing the observed data and imputed data so that it can help determine the independent variables for the credit risk modelling in subsequent. The second section provides the result and discussion of cross-section analysis (Logistic regression and Shrinkage regression model), and non-linear effect (GAMs model) will be discussed as well. The third section presents the panel data models analysis in terms of how to select macroeconomic variables and discuss their impact on SMEs' performance. The final section provides a conclusion of the results of this thesis.

Chapter six Conclusion:

This chapter provides a conclusion of the thesis and answers the research questions. This chapter also illustrates the implication of the research and points out the limitation of this thesis and possible future improvement for missing data handling, biased data, credit risk modelling, and a smaller subset of SMEs.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

The literature associated with missing data, relevant regulations (Basel Accord and IFRS 9) and credit risk models will be reviewed in this chapter. The problem of the missing data is unavoidable in practice since several variables reply on the SME supplying relevant information. Removing the incomplete observations before performing statistical analysis was preferred as it was easy to operate. Researcher, though, suggested alternative approaches such as maximum likelihood (ML) and multiple imputation (MI) appears. Yet, the mechanism of missing data needs to be understood before one applies any other technique to deal with missing variables. With the increasing computational power, these methods have changed researchers' behaviour to deal with missing data. Hence, there has been greater effort to reveal the 'truth' behind the data, and this is beneficial for the building of credit risk models. Section 2.3 reviews the mechanism of missing data and the two methods used.

Basel Accord II has been central to bank's regulations. It is believed that banks are able to develop steadily under the capital requirement of credit risk, market risk as well as operation risk. Yet, the occurrence of the credit crisis destroyed this belief. There was a gloomy picture of an economy sliding into recession in the whole world between 2008 and 2009. In order to reduce the possibility of similar events in the futures, Basel III and International Financial Report Standard (IFRS) 9 were released, that increased the capital requirement and introduction of additional buffers by Basel III, and a change to incurred loss model from expected credit loss (ECL) by IFRS 9 is regarded as remedial measures. These will be discussed in the later sections. Section 2.3 and section 2.4 reviews the history of Basel Accord and the impairment model under IFRS 9, respectively. Section 2.5 discussed the definition under different standards. Section 2.6 compares these two regulations. Section 2.8 provides

the definition of ‘credit’, as well as the objective of credit risk modelling. Then, Beaver (1966) and Edward I Altman (1968) approaches are introduced since they began the exploration of the credit models to predict business failure, and various methods targeted on SMEs credit modelling are also introduced. Finally, section 2.9 provides a summary and conclusion of this chapter.

2.2 Missing Data

Datasets are often partially observed in the real world. The general procedure of statistical analysis includes data collecting, model building and drawing conclusions from the model. Unfortunately, it always seems to be impossible to collect all the intended data especially when performing longitudinal or cross-sectional studies. In the era of big data, analysis requires a large amount of data and therefore there is a need to be able to utilise it appropriately.

It has long been a major concern in every field of study since missing data can have a significant effect on the derived model from the data (J. W. Graham, 2009). The problem of missing data is ubiquitous, yet the methodology for data analysis often assumes that the dataset is complete. Researchers have attempted to solve the problem by deletion, imputation, regression and use of dummy variables. However, a relatively strict assumption about the reason for the missing data is required before it is appropriate to apply this technique otherwise it is likely to produce biased estimates (Little and Rubin, 1987). Among these methods, list-wise deletion that removes cases with missing data is the most popular, which enables researchers to make statistical inferences directly from a “complete” dataset (A. M. Wood, White, & Thompson, 2004; Jelicic, Phelps, & Lerner, 2009; Peugh and Enders, 2016). Wilkinson (1999) has warned that list-wise deletion is the worst methods available for practical applications. It would be no exaggeration to say that a relatively small reduction of observations causes a significant decrease in the valid sample sizes, which affects the the probability that the test will reject the null hypothesis when it is false and it develops a biased estimation of parameters; consequently, it may

make the analysis of the study more complex and lead to invalid conclusions (Little and Rubin, 1987; Little, 1992).

Missing data obscures the true values for which a meaningful analysis is required (Little and Rubin, 1987). Neglecting the missing data problem can result in adverse consequences such as the loss of statistical power of a given analysis due to the reduction of the sample size, or even worse, missing values may invalidate the conclusions for the data and lead to wrong statistical inference. Today, disadvantages of these methods are well known in both the statistical and applied literature (J. W. Graham, 2009; Little and Rubin, 2014).

2.2.1 Missing Data Mechanisms

The performance of missing data techniques strongly depends on the mechanism that generated the missing values. Donald B Rubin (1976) established a theoretical framework for missing data problem in the form of three mechanisms, which bases on the probability of missingness, namely: missing complete at random (MCAR), missing at random (MAR), and missing not at random (MNAR) respectively. The mechanism can be interpreted as a probability distribution for the missing data. The result of this evaluation is significant since it limits the possible approaches of dealing with the missing data in further analysis. The crucial role of the mechanisms in the analysis of data with missing values was largely ignored until the concept was formalized in the way of treating the missing data indicators as random variables and assigning them a distribution (Little and Rubin, 2014).

Let Y = (y_{ij}) denote an (n×T) data matrix without missing values, with i^{th} row y_i = (y_{i1},…, y_{iT}) where y_{ij} is the value of variable Y_j for observation i. Donald B Rubin (1976) proposed that missingness is a variable that has a probability distribution, and defined missing data indicator matrix $R = r_{ij}$, such that $r_{ij} = 1$ if y_{ij} is missing and $r_{ij} = 0$ if y_{ij} is not missing, and this R matrix has the same size as the data matrix (matrix Y). R could be analysed by researchers

according to the probability models. For the purpose to further explanation, the notations are used: P() is a generic symbol for a probability distribution, and R is the missing data indicator.

The complete data (Y_{com}) can be divided into an observed (Y_{obs}) and a missing part (Y_{mis}):

$$Y_{com} = Y_{obs} + Y_{mis}, \qquad (Y_{obs} \cap Y_{mis} = \emptyset) \tag{1}$$

where ∅ are a set of unknown parameters that describe the relationship between R and the data.

If the probability of missing data on a variable Y depends only on the component Y_{obs} but not on Y_{mis}, that is, if

$$P(R \mid Y_{com}) = P(R \mid Y_{obs}, \emptyset) \tag{2}$$

the data are defined as MAR. In other words, above equation means that the probability to R takes on a value of zero or one can depend on Y_{obs} (Joseph L. Schafer and Graham, 2002).

MCAR is a stronger assumption than MAR, which is that the probability of missing data on a variable Y neither depends on Y_{obs} nor Y_{mis}. Missingness is completely unrelated to the data. That is

$$P(R \mid Y_{com}) = P(R \mid \emptyset) \tag{3}$$

The probability of MCAR data is constant, and MCAR data is not deemed common in real life because missing values are most likely dependent on other variables. Under this assumption, it is possible to train models using deletion method without bias, but this is not recommended because of information loss. Furthermore, incorrect specification of the MCAR assumption can lead to bias. As a result, most imputation methods are not based on the MCAR mechanism.

Finally, the mechanism is called MNAR if the probability of missing data on a variable Y can depend on other variables (i.e., Y_{obs}) as well as on the

unobserved underlying values of Y itself (i.e., Y_{mis}) (Craig K Enders, 2010). That is

$$P(R \mid Y_{com}) = P(R \mid Y_{obs}, Y_{mis}, \emptyset) \quad (4)$$

The MNAR assumption can be problematic to work with as the factors that influence Y_{mis} are difficult to study.

MAR is the most commonly used assumption in imputation methods as it does not carry the risk of MCAR misspecification and the complexity of MNAR (Craig K Enders, 2010). One significant concept regarding missing values is related to the pattern of missing data. If a hierarchy of missing values could be observed within the data matrix, so that observing a particular variable X_b for a subject implies that X_a is observed, for a < b, then the missing value is said to be monotone pattern. There is a special method for monotone pattern missing values. In general, most of the cases belongs arbitrary pattern (Horton and Kleinman, 2007).

Missing data theory (Donald B Rubin, 1976) involves two sets of parameters: the parameters that have no missing data and the parameters that describe the probability of missing data (i.e., ø). Researchers rarely know why the data are missing, so it is impossible to determine or estimate ø with any certainty, but ø may influence the estimation of the parameters of interest although ø have no substantive value. Rubin's missing data theory is important because he clarified the conditions that need to exist in order to accurately estimate the parameters of interest without also knowing the parameters of the missing data distribution (i.e., ø). Rubin showed that likelihood-based analyses such Maximum Likelihood Estimation (ML) and Multiple Imputation (MI) do not require information about ø if the data are MCAR or MAR (Little and Rubin, 1987; Donald B Rubin, 1987a). For this reason, the missing data literature often describes the MAR or MCAR mechanisms as ignorable missingness because there is no need to estimate the parameters of the missing data distribution when performing analyses. Careful consideration of the missing data mechanism is important because different types of missing data require

different treatments (Joseph L. Schafer, 2003; Paul D. Allison, 2016). In practical terms, adopting a MAR-based approach such as MI, ensures obtaining accurate estimates in a broader range of circumstances than simply removing incomplete cases. Importantly, this advantage is largely unrelated to the amount of missing data; if the imputation procedure satisfies MAR, the resulting estimates can tolerate extreme levels of missingness (e.g., 50% or more) (Craig K Enders, 2010).

In addition, MCAR is the only missing data mechanism that can be tested. One of the most common methods is to use a series of independent t-test to compare missing data subgroups (Craig K Enders, 2010). This approach separates the missing and the complete cases for a particular variable and uses a t-test to examine group mean differences on other variables in the dataset. The MCAR mechanism implies that the cases with observed data should be the same as the cases with missing values on average. Consequently, a non-significant t-test provides evidence that the data are MCAR. By definition the presence or absence of MNAR can never be demonstrated using only the observed data. Thus, without additional information, it is impossible to test whether MAR or MNAR holds. Nevertheless, even if MAR was assumed erroneously, there are indications that departures from it do not necessarily cause serious consequences (Joseph L. Schafer and Graham, 2002).

In reality, it is not common to observe data belonging to MCAR. Similar situations can be found in the credit risk datasets, collecting data from financial statement based on certain reporting regulations. Firm size is one of the factors that determines the quantity of financial data which have to be reported under UK financial reporting standards (Gov, 2018). For example, small companies can choose to disclose less information than medium-sized and large companies. Therefore, it is likely that a smaller company does not report certain balance sheet positions which can be found for larger companies. Therefore, the possibility of missing value depends on the size of the company.

Firm size is reflected in balance sheet positions such as net sales which in turn is a frequent component in common financial ratios. Therefore, it is plausible to assume that the probability of missing a certain financial ratio depends on the value of other financial ratios. In other words, these considerations violate the concept of missing values in the financial statement data as MCAR. On the other hand, there is no reason to assume that missing data in financial statements depend on their real value, i.e. that they are MNAR.

From a practical standpoint, missing data mechanisms are essentially assumptions that govern the performance of different analytic techniques and dictate the accuracy of a missing data handling procedure. Traditional methods assume an MCAR mechanism with few exceptions and will yield biased parameter estimate when data are MAR or MNAR. Although there is loss of statistical power, MI and ML yield unbiased parameter estimates with MCAR or MAR data (J. W. Graham, 2009; Craig K Enders, 2010), but they are still not perfect as they will produce bias with MNAR data. Methodologists have developed analysis methods for MNAR, but these approaches require strict assumptions that limit their practical utility (Craig K Enders, 2011).

Therefore, based on the above considerations, it is reasonable to assume that the mechanism of missingness is MAR so that mechanism of missingness is ignorable and it is available to apply MAR-based missing data handling method (e.g., ML and MI).

2.2.2 Advanced Methods for Missing Data

A revolution of dealing with missing data began with Donald B Rubin (1976) missing data mechanism, which is one of the most influential articles developing a theoretical framework for missing data. It attempts to define the reasons of ‘missingness’, and this framework still remains in use currently. Later in 1987, two major books (Little and Rubin, 1987; Donald B Rubin, 1987a) were published, and laid the foundation of missing data analysis methods for the next few decades, because these books introduced two methods for

missing data: multiple imputation (MI) and maximum likelihood estimation (ML), which enable producing unbiased estimates of the parameters and provide an estimate of the uncertainty about those estimates.

These two primary schools for dealing with missing values introduce an advanced and practical way of how to deal with the missing data problem. On one side, there are model-based methods mainly built around the formulation of the Expectation-Maximization (EM) algorithm made popular by (Dempster, Laird, & Rubin, 1977). This technique makes the computation of ML estimator feasible in problems affected by missing data. In short, the EM algorithm is an iterative procedure that produces ML estimates. The idea is to treat the missing data as random variables to be removed by integration from the log-likelihood function as if they were not sampled. The EM algorithm allows dealing with the missing data and parameter estimation in the same step. The major drawback of this model-based method is the requirement of the explicit modelling of joint multivariate distributions and, thus, tend to be limited to variables deemed to be of substantive relevance (John W Graham, Cumsille, & Elek‐Fisk, 2003). For example, in a regression analysis, the ML estimates are coefficients that minimise the sum of the squared standardised distances between the observed data and the regression line. Some methodologists have characterised ML estimation as implicit imputation because it does not produce a complete dataset (Widaman, 2006). Rather, the procedure uses all the available data to estimate a specific set of model parameters and their standard errors.

Furthermore, this approach requires the correct specification of usually high-dimensional distributions, even of aspects which have never been the focus of empirical research and for which justification is hardly available. According to J. W. Graham (2009), the parameter estimators (means, variances, and covariance) from the EM algorithm are preferable over a wide range of possible estimators, based on the fact that they enjoy the properties of maximum likelihood estimation.

The second approach deals with model-based missing data procedures and was introduced by (Donald B Rubin, 1987a) with his concept of MI. Instead of removing the missing values by integration as EM does, MI creates several versions of a dataset, each of which contains different estimates of the missing values. MI uses a regression model to fill in the data, treating incomplete variables as outcomes and complete variables as predictors. To avoid imputations based on a single set of regression parameters, an iterative algorithm uses Bayesian estimation to update the regression model parameters, and it uses new estimates to generate each set of imputations. The substituted values are called “imputed” values, hence the term “Multiple Imputation.” Having generated a set of ‘filled-in’ data, the researcher then performs one or more statistical analyses on each complete dataset to obtain imputation-specific estimates and standard errors. The final step is to pool coefficient estimates and standard errors into a single set of results. In addition, MI separates the solution of missing data problem from the solution of the complete data problem (Stef Van Buuren, 2012). Thus, when talking about model, it refers to two kinds of model: imputation model and analysis model. MI inferences assume that the analyst’s model is the same as the imputer’s model. Yet in practice, it can be accepted as long as variables in analysis model are a subset of variables in imputation model. The missing data problem (imputation model) is solved first, then the complete data problem (analysis model).

MI can be summarized in three steps. The first step is to create m sets of completed data by replacing each missing value with m imputed values. The second phase consists of using standard statistical methods for separate analysis of each completed dataset as if it were a “real” completely observed dataset. The third step is the pooling step where the results from m analyses are combined to form the final results and allows statistical inference in the usual way. This technique has become one of the most advocated methods

for handling missing data. A schematic overview of the MI procedure is depicted in Figure 2-1 for a three-time imputation.

The MI framework comprises three models: the complete data model, the nonresponse model, and the imputation model. The complete data model is the one used to make inferences of scientific interest. For example, a linear regression including the outcome and explanatory variables of an experiment. The nonresponse model represents the process that leads to missing data. The covariates in the nonresponse model are not primarily of interest, and they are not necessarily part of the complete data model. The imputation model is the model from which plausible values for each missing datum are generated. A problematic step of MI procedures is the specification of the imputation model because the validity of the analysis of the complete data model strongly depends on how imputations are created. If the imputation model is not correctly specified, then final inferences may be invalid.

One of the most critical assumptions for both methods is that a joint distribution of all variables in the dataset including outcome variable is multivariate normal distributed (Pigott, 2001). This assumption seems to exclude the use of categorical variables, but Joseph L Schafer (1997) discussed how normal distribution can be relaxed as long as the categorical variables are complete observed; otherwise, multivariate normal assumption would be inappropriate if categorical variables in the dataset have high rates of missing observations (Pigott, 2001). On the other hand, with normally distributed data, a common set of input variables and a sufficiently large sample size, there is no theoretical reason to expect differences between ML estimation and MI (Joseph L. Schafer, 2003). Empirical studies suggested that the two methods usually yield similar estimates and standard errors (Collins, Schafer, & Kam, 2001). P. Allison (2012a) indicated that ML is preferable as its simplicity, and it produces a deterministic result while the performance of MI is unstable, which means various adjustments of input have to be used. Variables used, numbers of imputation, and other factors have impacts on the accuracy of parameter

estimates. Credit scoring datasets often feature complexities that include a mix of categorical and continuous, even semi-continuous variables.

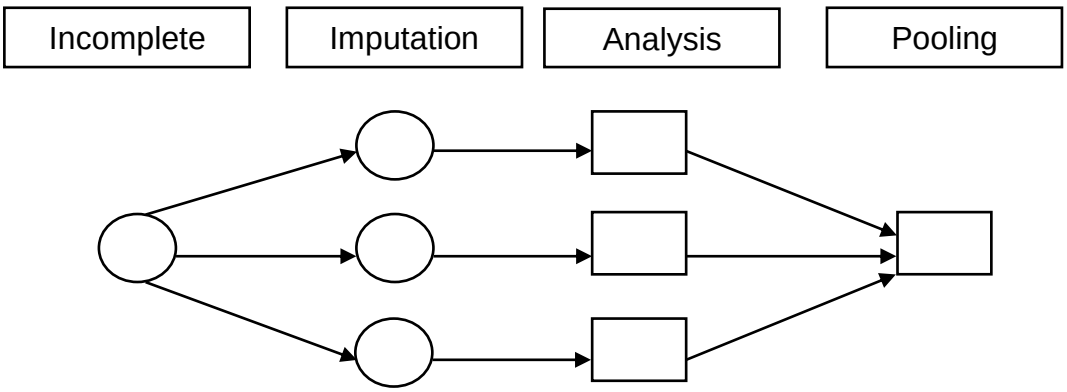


Figure 2-1 Overview of the MI procedure

There are two ways of specifying imputation models: Joint modelling (JM) and fully conditional specification (FCS). Joint modelling involves specifying a multivariate distribution for the variables whose values have not been observed conditional on the observed data and then drawing imputations from this conditional distribution by Markov chain Monte Carlo (MCMC) techniques (Joseph L Schafer, 1997).

Within the JM framework, researchers have developed imputation procedures for multivariate continuous, categorical and mixed continuous and categorical data based on the multivariate normal, log-linear and general location model, respectively (Donald B Rubin, 1987a; Joseph L Schafer, 1997; Little and Rubin, 2014). This methodology is obviously attractive as valid imputations may be generate by linear regression equations. However, because of the normality and linearity, it may not be well suited for imputing categorical variables as well. Besides, the use of joint modelling strategy can be challenging for large datasets with hundreds of variables of varying types (Azur, Stuart, Frangakis, & Leaf, 2011).

On the other hand, with the fully conditional specification, also known as multivariate imputation by chained equations (MICE) (Buuren and Groothuis-

Oudshoorn, 2010), a univariate imputation model is specified for each variable with missing conditional on other variables of the dataset. Initial missing values are imputed with a bootstrap sample, and then subsequent imputations are drawn by iterating over conditional densities (S. van Buuren, 2007; Buuren and Groothuis-Oudshoorn, 2010). MICE is attractive as an alternative to joint modelling in cases where no suitable multivariate distribution can be found. MICE is thus considered as a better tool in credit scoring because it gives researchers the flexibility to tailor the missing data handling procedure to match a particular set of analysis goals. Lazure (2017) conducted a credit classification with missing data of a simulation study on German credit data. The author applied random forest, support vector machine, MICE and predictive mean matching to handle missing data, and concluded that MICE in tandem with predictive mean matching can be an optimal method of imputation method in the field of credit scoring.

Since ML and MI appeared in 1987, there has been a substantial increase in missing data research about these two methods. Joseph L Schafer (1997) developed various joint modelling technique for imputation under the multivariate normal, the log-linear, and the general location model. J. L. Schafer (1999) answered questions about how to apply MI in reality. Raghunathan, Lepkowski, Van Hoewyk, & Solenberger (2001) applied MICE on a relatively complex data structure, involving continuous, categorical, counts, and mixed variables. K. J. Lee and Carlin (2010) compared joint modelling and MICE, and concluded that both they can provide similar result in a standard regression including different scaled variables. Besides, this article recommended that in order to avoid biases and produce a reliable estimation, it is necessary to transform skewed variables to a symmetric distribution. White, Royston, & Wood (2011) provided broader issues and guidance for practice in the research area of mental health with Stata code fragments, but it also indicated disadvantages of MICE. More recently, C. K. Enders (2017) described numeric practical issues that clinical researchers are

likely to encounter when applying MI. Florez-Lopez (2017) showed that MI models are able to provide satisfactory solutions in credit risk missing data.

Although reporting practices have improved, the application of MI and ML on missing data handling techniques is far from uniform because traditional techniques are always simple to apply, and mainly because there was a lack of software option for MI and ML. A number of researchers still heavily relied on old methods handling missing data (T. E. Bodner, 2006; Peugh and Enders, 2016). Until the late 1990s, MI and ML became available in statistical software packages. Joseph L Schafer (1997) published the first widely available general-purpose imputation algorithm and made it available in “NORM” package. This package is now widely available within a number of statistical packages, such as SPSS. At the beginning of the 2010s, Journal of Statistical Software published three papers about the application of MI based on three different main statistical software: R, SAS, and Stata (Buuren and Groothuis-Oudshoorn, 2010; Royston and White, 2011; Yuan, 2011).These papers provided a comprehensive guideline of how to use MI on each software. On the other hand, an increasing number of people attached more importance to solve missing data by advanced approaches.

In addition, one can use non-parametric methods to generate imputations, like hot deck methods. Based on hot deck methods, the missing values are imputed by finding a similar but observed unit, whose value serves as a donor for the record of the similar but incompletely observed unit. The most popular are k-nearest neighbour algorithms (KNN) from which the best known method for generating hot-deck imputations is the Predictive Mean Matching (PMM) (Little, 1988), which imputes missing values employing the nearest-neighbour donor distance base on expected values of the missing variables conditional on observed covariates. There are several advantages of KNN imputation. It is a simple method that seems to avoid strong parametric assumptions, it can easily be applied to various types of variables to be imputed, and only available and observed values are imputed (Schenker and Taylor, 1996; Andridge and

Little, 2010). However, the final goal of the complete data statistical analysis is to make inferences about the population represented by the sample; therefore, the plausibility of imputed values is not the defining factor in choosing an imputation model over another. Instead, the proper criterion is the validity of the final analysis of scientific interest (Salfrán Vaquero, 2018).

2.3 Credit Risk Models

2.3.1 Definition

The term ‘credit’ refers to an amount of money that is loaned to a consumer by a financial institution and which must be repaid with interest in instalments (Hand and Henley, 1997). Traditionally, the financial credit risk indicates the risk associated with financing. Specifically, credit risk has been defined as the likelihood that a borrower will be unable to repay the loan according to its term (Golin and Delhaise, 2013); or the risk of losses due to an increased probability of default or default of a debtor from mark-to-market aspect (Nehrebecka and Polski, 2016). Besides, Basel Committee explains a default event on a debt obligation in the two following ways:

- It is unlikely that the obligor will be able to repay its debt to the bank without giving up any pledged collateral
- The obligor is more than 90 days past due on a material credit obligation.

Accordingly, given several companies labelled as bad/good credit or bankrupt/healthy with a set of financial variables that describe the situation of a company over a given period, financial credit risk assessment aims to solve the problem stated as follows:

- Predicting the probability that a company if belongs to which risk group (high-risk group or low-risk group) during the following years
- Predicting if the company is going bankrupt.

The former problem is called credit scoring, and the latter problem is called bankruptcy (failure) prediction. Credit scoring is the name used to describe a more formal process of determining how likely applicants are to default with their repayments. Both are solved similarly as a binary classification task. The

performance of those tasks solved by banking institutions has a crucial impact on the development of a country's economic, which was evidenced by the last financial crisis. There have been vast literatures on solving the problem since a reasonable prediction provides an early warning about any possible problems regarding cash flow and offers a chance to react quickly (Nehrebecka and Polski, 2016).

2.3.2 Business Failure Prediction

Corporate credit models originated from the work of bankruptcy prediction in the 1960s, and the pioneers of the credit scoring approach are Beaver (1966), and Edward I Altman (1968).

Beaver (1966) was one of the first researchers to study the prediction of bankruptcy using linear discriminant analysis (LDA) with financial statement data. It is based on studying one financial ratio at a time and on developing a cut-off threshold for each ratio. More specifically, he finds that the cash flow to total debt ratio shows the best performance in minimizing the number of errors made in classifying firms as failed and non-failed. However, his analysis is very simple in a univariate analysis without considering the combing influences of all indicators.

Edward I Altman (1968) developed a Z-score model with the classical multivariate discriminant analysis technique (MDA). It is based on applying the Bayes classification procedure, under the assumption that the two classes have Gaussian distributions with equal covariance matrices. The covariance matrix and the class means are estimated from the training set. The model incorporates the following financial ratios as inputs, and these particular financial ratios have been widely used even for other non-linear models:

- X1: working capital/total assets;
- X2: retained earnings/total assets;
- X3: earnings before interest and taxes/total assets;

- X4: market capitalization/total debt;
- X5: sales/total assets.

$$Z = 1.21x_1 + 1.40x_2 + 3.30x_4 + 0.604x_4 + 0.999x_5 \quad (5)$$

The firm is less risky if its Z-score is greater than the cut-off. The discriminant threshold (cut-off value) used to distinguish predicted defaulting from predicted performing companies is fixed at z = 2.675. After that Discriminant Analysis was widely used and discussed (Deakin, 1972; Blum, 1974; Edward I Altman, Haldeman, & Narayanan, 1977; Eisenbeis, 1977; Taffler, 1982; Lo, 1986; Edward I. Altman, Marco, & Varetto, 1994; Mika, Ratsch, Weston, Scholkopf, & Mullers, 1999).

Wilcox (1971) further extended Beaver's work to apply 'gambler's ruin theory' to business failures using accounting data to offer an explanation empirical result. The author provided an intrinsically probabilistic approach applied to corporate default description and proposed the 'time to default' concept for the first time.

Then, the rise of conditional probability regression models represented by logit Ohlson (1980) and probit (Zmijewski, 1984) was because Altman's Z-score model and conditional probability regression models are essentially linear models that classify between healthy/bankrupt firms using financial ratios as inputs, but the latter can calculate the probability of default in a predefined time period. The logistic regression approach is a linear model with a sigmoid function at the output. Since the output is in between 0 and 1, the model has a simple probabilistic interpretation. Since then, logistic regression is so dominant that many studies (Coughlan and Schmidt, 1985; Gilbert, Menon, & Schwartz, 1990; Hua, Wang, Xu, Zhang, & Liang, 2007; S. Y. Kim, 2011; Inam, Inam, Mian, Sheikh, & Awan, 2018) have chosen it to be the comparative tool and industry benchmark to quantify the risk of bankruptcy or financial distress of companies

In addition, Expert-based approach of deciding whether to grant credit to a particular individual use human judgment of the risk of default, based on experience of previous decisions. However, economic pressures resulting from increased demand for credit, allied with greater commercial competition and the emergence of new computer technology, have led to the development of sophisticated statistical models to aid the credit granting decision. There are two main approaches to loan default/bankruptcy prediction.

The structural approach is modelling the underlying dynamics of interest rates and firm characteristics and deriving the default probability based on economic and financial theoretical assumptions Atiya (2001). This approach tries to model an estimate of the formal relationships that associated the relevant variables of the theoretical model. Yet, structural form models do not apply to (unlisted) SMEs PD modelling since these models require market data, value of the SME, (Modina and Pietrovito, 2014), which often either does not exist or the stock might only be irregularly traded and may be misleading (Sohn, Kim, & Moon, 2007; Rikkers and Thibeault, 2009). The outstanding representative was the asset-based structural model by Merton (1974) who transplanted the option pricing mechanism (BS model) in detecting bankruptcy and thereafter the Merton-typed structured models are also very practised in many practices and developed further by (Longstaff and Schwartz, 1995). This model views a firm's equity as a call option on the firm (held by the shareholders) to either repay the debt of the firm when it is due, or abandon the firm without paying the obligations (De Laurentis, Maino, & Molteni, 2011). Following the Merton approach and applying Black Scholes Merton formula, the default probability is consequently given by:

$$\mathrm{PD} = \mathrm{N}\left(\frac{\log\left(\frac{\mathrm{F}}{\mathrm{V_A}}\right) + (\mu - 0.5{\sigma_\mathrm{A}}^2)\mathrm{T}}{\sigma_\mathrm{A}\sqrt{\mathrm{T}}}\right) \qquad (6)$$

where F is the debt face value, $(\mathrm{V_A})$ is the firm's asset value (equal to the market value of equity and net debt), μ is the 'risky world' expected return, T is the remaining time to maturity, σ_A is the asset value volatility, N(.) is the cumulated normal distribution operator. The probability of default can be

derived by modelling the market value of the firm as a geometric Brownian motion. What makes this model successful is its reliance on the equity market as an indicator, since it can be argued that the market capitalization of the firm (together with the firm's liabilities) reflect the solvency of the firm.

Reduced-form approach is the statistical-based approach, in which the final solution is reached using the most statistically suitable set of variables and disregarding the theoretical and conceptual causal relations among them (De Laurentis, et al., 2011). Specially, instead of modelling the relationship of default with the characteristics of a firm, this relationship is learned from the data. This approach uses predictor variables from application forms and other sources to yield estimates of the probabilities of defaulting. An acceptance or rejection decision is taken by comparing the estimated probability of defaulting with a suitable threshold, such as 0.5. Standard statistical methods used in the industry are discriminant analysis, linear regression, logistic regression and decision trees, etc.

In addition, the modern credit risk measurement model includes four major approaches: KMV model, CreditMetrics, Credit Portfolio View, and CreditRisk+. Merton's model has been successfully developed into a successful commercial product by KMV Corporation. The J.P. Morgan's CreditMetrics (Morgan, 1997) and McKinsey's Credit Portfolio View (Wu and Olson, 2010) are directly related to the credit rating mechanism. The CreditRisk+ product (Suisse, 1997), developed by Credit Suisse Financial Products, is based on the same concept of modelling default as a Poisson process. Before CreditRisk+, Jarrow and Turnbull (1992) modelled default as a point process, where the time-varying hazard function for each credit class is estimated from the credit spreads. Besides, the heuristic and numeric approach associated with artificial intelligence and machine learning are rapidly developing recently. S.-M. Lin (2007b) presented a comprehensive literature review on credit risk modelling from different aspects, including market-based models, accounting-based models, machine learning, expert system, and portfolio credit risk models.

2.3.3 Credit Risk Models for SMEs

SMEs are the subset of the whole population of all types of corporations but from the growth point of view, SMEs may be the early stage of a large corporations - all large companies grow from small ones when they were firstly incorporated. They do have difference in many aspects which need to be addressed in the risk models. E. I. Altman and Sabato (2007) also suggested that from a credit risk point of view, SMEs are different from large corporations for a number of reasons. The following section will review some models with special focuses on SMEs.

The first study of SME credit models may come from (Edmister, 1972) who employed MDA in distinguishing failed small businesses from non-failed ones. His sample consisted of 42 companies over 1954 to 1969 and initially 19 ratios were tested but only seven of them were left in the final equation of discriminant analysis. His pioneering work is rather limited because of the small sample size and biased selection of samples – he only included those companies with at least three-year annual reports. So his model is inapplicable on start-up SMEs but in the first three years of start-up, the mortality rate is much higher than what happens next (Bruderl, Preisendorfer, & Ziegler, 1992). His final equation took categorical values rather than real values of ratios based on Beaver (1966) but was in the form of Edward I Altman (1968)'s structure which lost much information in the transformation.

To study how to specifically model credit risk for SMEs, Table 2-1 lists literature review that various approaches are applied to study credit risk in different countries. Majority of researchers select statistical approaches to estimate credit risk.

As this research focuses on SMEs loans which have a high number of applications and each loan size is relatively small, especially due to special treatment of SMEs in Basel II, more SMEs are treated as retail obligors and

scorecard models are widely used for the SMEs segment. In addition, Merton type models are usually developed by efficient market theory under which it is assumed that a listed company's asset value could be fully represented by their share prices. Therefore, Merton type models may face restrictions when applied to a large SMEs portfolio.

Data limitations restrict the modelling choices. The wholesale commercial loans models use rich information concerning companies' financial health which comes from rating agencies and financial markets prices. In general, this information is available in the form of time series. It allows to assess the long run stability of the main building blocks of any credit risk model. It also allows analytically determination of the probability distribution of potential losses or to proceed to historical simulations (Dietsch and Petey, 2002). While in the SMEs case, the majority are not listed companies, rating-based approaches like CreditMetrics, Credit Portfolio View, and CreditRisk+ are not practical, and even if they are, trading of their shares may not be active as large corporations in the stock market, thus it is inappropriate to apply efficient market theory to analyse. E. I. Altman and Sabato (2007) suggested that banks should develop credit risk models specifically addressed to SMEs in order to minimize their expected and unexpected losses.

To conclude, this section reviews some credit scoring models. Statistical-based models generally can provide better understanding on the explanatory variables by estimating their parameters. For retail loans, the most widely used model for PD estimation is logistic regression, which is a statistic model. Logistic models could provide direct estimation of obligor's PD, give clear explanation of rejections and do not have high demand on data and hardware. This research seeks to develop the statistic models in several different ways. Firstly, logistic models' performance is tested by SME crisis data. Second step is to extend logistic models by adding non-parametric smoothers which captures non-linear trends of SMEs during the 'credit crunch'.

Table 2-1 Summary of credit risk models for SMEs

Dietsch and Petey (2002)	Authors proposed an internal credit risk model for SMEs loans, which enables us to compute the value at risk based on French SMEs.
Edward I Altman and Sabato (2005)	Authors developed a one-year default prediction model using a logit model over 1994-2002 based on U.S. SMEs.
Behr and Güttler (2007)	Authors studied a scoring model using a binary logistic regression model based on German SMEs.
Ze-jing and Fu-qiang (2008)	Authors studied credit risk using the KMV model with tunable parameters based on listed SMEs in China.
Fantazzini and Figini (2009)	Authors proposed a non-parametric approach to study credit risk based on Random Survival Forests (RSF) and compare its performance with a standard logit model using SMEs data in Germany.
Gupta, Wilson, Gregoriou, & Healy (2014)	Authors modelled one-year default risk of domestic and international UK SMEs using a dynamic logistic regression technique with a similar set of explanatory variables between 2000 and 2009.
Li, Niskanen, Kolehmainen, & Niskanen (2016)	Authors used the data of Finnish SMEs from the fiscal year 2004 to 2012 to estimate credit risk based on a hybrid model which combines the logistic regression approach and artificial neural network (ANN).
Gupta, Gregoriou, & Ebrahimi (2017)	Authors compared the discrete-time hazard models and the continuous-time Cox Proportional Hazards model in predicting bankruptcy and financial distress of the USA SMEs.

The choice of the model, however, would depend on the circumstances. Among all the methods surveyed here, there is no single model which may be

termed as a standard solution that would suit all banks. A variety of factors determine the best fit for the purpose.

2.4 Framework of Basel Accord

Banks have a vital function in the economy. They have access to funds through collecting savers' money, issuing debt securities, or borrowing on the inter-bank markets. The funds collected are invested in short-term and long-term risky assets, which consist mainly of credits to various economic actors. Through centralizing any money surplus and injecting it back into the economy, large banks are the heart maintaining the blood supply of our modern capitalist societies. In addition, systemic risk is a concern to the bank sectors. It is the risk that a failure of a large bank will lead to failures by other large banks and so a collapse of the financial system. So, it is no surprise that they are subject to regulatory constraints, without there would be dangers for the global economy (Balthazar, 2006).

In response to a significant liquidation of a Europe-based bank in June 1974, the Basel Committee on Bank Supervision (BCBS) was founded in late 1974, as an international forum where members could cooperate on banking supervision matters under the direction and supervision of the Bank of International Settlements (BIS) in Basel, Switzerland. The BCBS aims to enhance "financial stability by improving supervisory know-how and the quality of banking supervision worldwide." This is done through regulations known as accords. Specifically, the main purpose of bank regulation is to ensure that a bank keeps enough capital for the risks it takes since it is not possible to eliminate the possibility of a bank failing, but governments want to make the probability of default for any given bank very small. By doing this, members hope to create a stable economic environment where private individuals and businesses have confidence in the banking system (Hull, 2012). As time goes by, the evolution process of international bank regulation (Basel Accord) is

shown in Table 2-2 in order to follow up the complex and changeable globally financial environments (Dionne, 2013).

Table 2-2 History of Basel Accord

Time	Implementation	Decision
1988	1992	Basel I Accord (start of credit risk)
1996	1998	1996 Amendment (start of market risk)
2004	2007	Basel II Accord (credit risk reform, and start of operational risk)
	2011	Basel 2.5
2010	2019 (fully)	Basel III Accord (start of liquidity risk, capital conservation buffer, countercyclical buffers)

When a bank (or other large financial institutions) gets into financial difficulties, governments have a difficult decision to make. If they allow the financial institution to fail, they are putting the financial system at risk. If they bail out the financial institution, they are sending the wrong signal to the market. There is a danger that large financial institutions will be less vigilant in controlling risks if they know that they are “too big to fail” and the government will always bail them out. During the market turmoil of 2008, the decision was taken to bail out some financial institutions in the United States and Europe. However, Lehman Brothers was allowed to fail in September 2008. Possibly, the United States government wanted to make it clear to the market that bailouts for large financial institutions were not automatic. However, the decision to let Lehman Brothers fail has been criticized because arguably it made the credit crisis worse (Sieczka, Sornette, & Holyst, 2011). Ivashina and Scharfstein (2010) further indicated that the failure of Lehman Brothers as a critical turning point in credit markets which resulted in an increased difficulty for banks to roll over short-term debt.

2.4.1 Basel I Accord

The Group of Ten (G-10)[7], the most industrialized countries, signed an accord in 1988 to supervise banks known as Basel I Accord, which was the beginning of international standard for bank regulation. The agreement obliges banks in member countries to hold a minimum amount of required capital to hedge against various risks and create a solvency reserve for the bank (Dionne, 2013).

Member countries can impose stronger regulations on their banks, strengthen the stability of international banking system, and set up a fair and a consistent international banking system in order to decrease competitive inequality among international banks, and to pave the way for a significant increase in Banks' commitment to risk measurement, understanding and management.

The key achievement of Basel I has been to define the bank capital and the bank capital ratio, also known as Cooke ratio[8]. The Basel I agreement formally defines capital based on two tiers: Tier 1 (Core capital) and Tier 2 (Supplementary Capital). Besides, capital requirement for the credit risk is defined as the proportion of risk weighted asset (RWA) of the bank. RWA is a bank's assets weighted according to risk. The total (credit) risk-weighted assets for a bank will be sum of its on- and off-balance sheet risk-weighted assets. A bank's assets weighted in relation to their relative credit risk (On-Balance Sheet Items) levels where there were totally four levels ranging from 0% to 100%, see Table 2-3.

[7] Belgium, Canada, France, Germany, Italy, Japan, Luxembourg, the Netherlands, Spain, Sweden, Switzerland, the United Kingdom, and the United States

[8] named after Peter Cooke from Bank of England

Table 2-3 Risk Weights for On-Balance Sheet Items

Risk Weight	Asset Category
0%	Cash, gold, claims on Organisation of Economic Co-operation and Development (OECD) countries such as US Treasury bonds and insured
20%	Claims on OECD banks and government agencies like US agency securities or municipal bonds
50%	Uninsured residential mortgages
100%	Loans to corporations, corporate bonds, claims on non-OECD banks

Zero weight is the most secure asset while 100 per cent weight is the riskiest asset. Accordingly, the Basel I required that capital to weighted risk assets should be set at 8% of which the Tier 1 capital will be at least 4%, and common equity in Tier 1 capital will be at least 2% (BCBS, 1988) i.e., the following inequalities must hold:

$$\text{Tier 1 ratio} = \frac{\text{Tier 1 Capital}}{\text{RWA}} \geq 4\% \tag{7}$$

$$\text{Total Capital ratio} = \frac{\text{Total Capital}}{\text{RWA}} = \frac{(\text{Tier 1 Capital} + \text{Tier 2 Capital})}{\text{RWA}} \geq 8\% \tag{8}$$

The Basel I Accord was heavily criticized as being too simple and somewhat arbitrary. It took into account the credit risk only but ignored the market risk and operational risk. It also had limited differentiation of credit risk with only four broad risk weightings. Besides, it ignored the different level of risks associated with different currencies and macroeconomic risk. For example, it assumes a common market to all factors though it is not true in reality where SMEs in developing countries and developed countries are allocated to the same risk weighting. On the other hand, it treated all corporate loans the same in terms of capital requirements. The creditworthiness of the borrower is

ignored. A loan to a corporation with an AAA credit rating is treated in the same way as one to a corporation with a B credit rating. In addition, Basel I was computed on the basis of book-value accounting measures of capital, not market values. Accounting practices could be different significantly across the G-10 countries and often produced results that differ markedly from market assessments. Also, in Basel I there was no model of default correlation (Hull, 2012). In conclusion, Basel I was considered too simplistic to address the activities of the complex banking institutions, but it was regarded as a meaningful beginning on banking supervision.

2.4.2 Basel II Accord

The above shortcomings led to a creation of a new Basel Capital Accord in 2004, known as Basel II (BCBS, 2006). The key innovation of Basel II is a "three pillars" concept:

- Pillar one – minimum capital requirements
- Pillar two – supervisory review
- Pillar three – market discipline

Among these three pillars, pillar one especially plays an important role as it introduces new approaches to determine and calculate capital requirements for credit risk[9]. The second pillar is associated with the supervisory review process. It includes both quantitative and qualitative aspects of the ways risk is managed in a bank. The third pillar requires banks to disclose more information about how they allocate capital and the risks they take.

Basel II divides the eligible regulatory capital of a bank into three tiers where the higher the tier, the less subordinated securities a bank is allowed to include in it. Each tier must have a certain minimum percentage of the total regulatory

[9] The capital requirement for market was introduced in 1996 Amendment using the method of Value at Risk (VaR). In addition, capital charge for operation risk was also considered in Basel II.

capital. The level of minimum capital requirement was continued to be maintained at 8% under the new framework[10]. In mathematics, that is:

$$\begin{aligned}\text{Minimum Total Capital} \\ = 0.08 \times (\text{Credit risk RWA} + \text{Market risk RWA} \\ + \text{Operational risk RWA})\end{aligned} \quad (9)$$

where RWA is risk weighted asset.

Basel II provided three different approaches to determine credit risk, namely standardized approach, foundation internal rating-based approach (F-IRB), and advanced rating-based approach (A-IRB).

2.4.2.1 Standardized approach

The standardized approach (SA) is similar to Basel I except for the distribution of risk weights. This approach considers the credit rating of assets in determining risk weights corresponding to various risk category based on ratings given by approved external credit rating agencies. The risk weights vary from 0% to 150% based on the risk category, and the higher the credit rating, the lower risk weight, see Table 2-4. The standard rule for retail lending (the maximum aggregated retail exposure to one counterpart cannot exceed an absolute threshold of €1 million.) is that a risk weight of 75% be applied[11]. When claims are secured by a residential mortgage or by commercial real estate, the risk weight is 35% or 100%, respectively[12]. According to Basel I, risk weights in retail and small business loans are placed in the 100% risk weight basket.

Clearly, it is not sufficiently sophisticated to use the standardized approach for large banks due to its simplicity. Standardized approach has increased risk sensitivity by considering expanded range of collateral, guarantees and credit

[10] (BCBS, 2006) paragraph 40.

[11] (BCBS, 2006) paragraph 69.

[12] (BCBS, 2006) paragraph 72, 74.

derivatives. Besides, the OECD status of a bank or a country is no longer considered important under Basel II. Yet, one loophole is that a bank can reduce its capital with unrated assets since unrated assets had lower risk weight than that of low rating assets. For example, the risk weight of an unrated bank is 50% while a bank rated below B- is 150%.

Table 2-4 Risk Weights as a Percent of Principal for Exposures to Countries, Banks, and Corporations Under Basel ll's Standardized Approach

	AAA to AA-	A+ to A-	BBB+ to BBB-	BB+ to BB-	B+ to B-	Below B-	Unrated
Sovereigns	0	20	50	100	100	150	100
Banks	20	50	50	100	100	150	50
Corporations	20	50	100	100	100	150	100

2.4.2.2 IRB-Approach

The IRB approach is based on the internal estimations made by the bank, which allows the bank to calculate capital requirements that are more sensitive to the risk.

Under the IRB approach, the capital requirement is based on Value at Risk (VaR) calculated over a one-year time horizon and a 99.9% confidence level. As shown in Figure 2-2, the idea behind IRB approach is to determine the unexpected loss (UL) since capital is used to compensate the UL. Hence, the capital required is VaR minus the expected loss (EL).

The VaR is calculated using the one-factor Gaussian copula model of time to default. Assume that a bank has a very large number of obligors and the i^{th} obligor has a one-year probability of default equal to PD_i. The copula correlation between each pair of obligors is R.

Worst-case default rate (WCDR) defines as the 99.9% quantile of the default rate distribution so that the bank is 99.9% certain it will not be exceeded next year for the i^{th} counterparty

$$\mathrm{WCDR_i} = \mathrm{N}[\frac{\mathrm{N^{-1}(PD_i)} + \sqrt{\rho}\mathrm{N^{-1}}(0.999)}{\sqrt{1-\rho}}] \quad (10)$$

where: N^{-1} (z) is the inverse cumulative distribution function (c.d.f.) for a standard normal random variable, i.e. the value of y such that N(y) = z. N(.) is the c.d.f. for a standard normal variable.

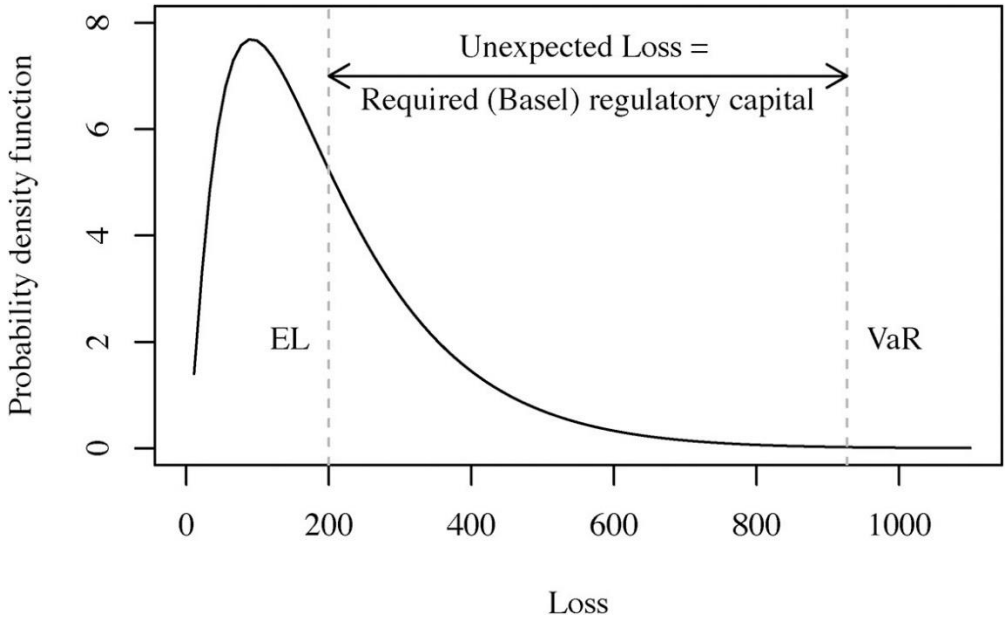


Figure 2-2 Capital requirement for IRB approach under Basel II taken from Krüger, Rösch, & Scheule (2018)

Basel II assumes that there is a relationship between ρ and PD, based on empirical research. In the case of corporate, sovereign, and bank exposure, the relationship is expressed as followed:

$$\begin{aligned}\mathrm{Correlation(R)} &= 0.12\frac{1-\mathrm{e}^{(-50\times\mathrm{PD})}}{1-\mathrm{e}^{-50}} + 0.24\left(1-\frac{1-\mathrm{e}^{(-50\times\mathrm{PD})}}{1-\mathrm{e}^{-50}}\right) \\ &\approx 0.12 + 0.12\mathrm{e}^{-50\times\mathrm{PD}}\end{aligned} \quad (11)$$

Following formulas are used to calculate capital requirement (Gordy, 2003):

$$\text{Capital} = \sum_{\text{i}} \text{EAD}_{\text{i}} \times \text{LGD}_{\text{i}} \times (\text{WCDR}_{\text{i}} - \text{PD}_{\text{i}}) \times \text{MA}_{\text{i}} \qquad (12)$$

where: Loss given default (LGD) measures the proportion of the exposure that will be lost if default occurs. Exposure at default (EAD) is estimated amount outstanding in a loan commitment if default occurs (in dollars). Probability of default (PD) measures the likelihood that the borrower will default over a given time horizon. Maturity adjustment (MA) is calculated as:

$$\text{MA} = \frac{1 + (\text{M} - 2.5) \times [0.11852 - 0.05478 \times \log(\text{PD})]^2}{1 - 1.5 \times [0.11852 - 0.05478 \times \log(\text{PD})]^2} \qquad (13)$$

where M is the maturity of exposure (Note that, when M = 1, MA is 1 and has no effect).The logic of the maturity adjustment is that if an instrument lasts longer than one year, there is a one-year credit exposure arising from a possible decline in the creditworthiness of the counterparty as well as from a possible default by the counterparty.

Under the F-IRB, banks only provide their own estimates of probability of default (PD) using historical data for past five years with a floor to 0.03% and rely on the supervisory estimates for other risk components: the loss given default, the exposure at default (EAD), and the effective maturity of the operation (M = 2.5 in most cases).

Regarding A-IRB, banks provide more of their own estimates of PD, LGD, EAD, and their own calculation of maturity with historical data for past seven years, subject to meeting minimum standards. Banks must always use the risk-weight functions provided in Basel Accords for the purpose of deriving capital requirements.

In the case of retail exposures, the model underlying the calculation of capital for retail exposures is similar to that underlying the calculation of corporate, sovereign, and banking exposures. However, the Foundation IRB and Advanced IRB approaches are merged and all banks using the IRB approach provide their own estimates of PD, EAD, and LGD. There is no maturity

adjustment, MA since it is believed that its maturity is shorter. The capital requirement is therefore

$$\text{Capital} = \sum_{i} \text{EAD}_i \times \text{LGD}_i \times (\text{WCDR}_i - \text{PD}_i) \tag{14}$$

The relationship between ρ and PD is:

$$\begin{aligned} \text{Correlation}(R) &= 0.03\frac{1-e^{(-35\times PD)}}{1-e^{-35}} + 0.16\left(1-\frac{1-e^{(-35\times PD)}}{1-e^{-35}}\right) \\ &\approx 0.03 + 0.13e^{-35\times PD} \end{aligned} \tag{15}$$

As PD increases, R decreases. The reason usually given for this inverse relationship is as follows. As a company becomes less creditworthy, its PD increases, and its probability of default becomes more idiosyncratic and less affected by overall market conditions. Correlations are assumed to be much lower for retail exposures than for corporate exposures, see Table 2-5.

Table 2-5 Relationship between PD and WCDR for firm, bank and retail exposure

		PD				
		0.1%	0.5%	1.0%	1.5%	2.0%
WCRD	Bank	3.4%	9.8%	14.0%	16.9%	19.0%
	Retail	2.1%	6.3%	9.1%	11.0%	12.3%

All in all, Basel II improves bank's flexibility to estimate capital requirement by the introduction of the internal ratings based approach (IRB) (S.-M. Lin, 2007a). The philosophy of capital formula is intended to be sufficient to cover unexpected losses over a one-year period that 99.9% certain will not be exceeded. The WCDR is the default rate that (theoretically) happens once every thousand years. The Basel committee reserved the right to apply a scaling factor (A typical scaling factor is 1.06[13]) to the result of the capital if it finds that the aggregate capital requirements are too high or low. Besides,

[13] (BCBS, 2006) paragraph 14.

default correlation is considered in Basel II since it is essential to determine the distribution of losses in a bank loan portfolio. Capturing the correlations between individual exposures is crucial in order to assess the risk at the portfolio level. In most of the credit risk models, the correlations measure the degree of sensitivity of PD to the systematic risk factors that represent the influence of the ‘‘state of the economy’’. Portfolio risk will be greater the more the bank loans tend to vary simultaneously in reaction to the realization of these risk factors. Hence, a crucial element in the estimation of loan loss distribution is a good calibration of parameters (PD).

The different risk weights and risk component estimates specified in Basel II thus lead to differential capital requirements for retail and corporate banking (Lim and Yong, 2017). Concerns have been raised that Basel Accord II will change the way banks analyse credits, introducing new credit risk management techniques and possibly reducing the lending activity toward SMEs. This is due to banks’ potential perception that SMEs carry higher risk and, hence, higher capital requirements than under Basel I (Edward I Altman and Sabato, 2005).

2.4.3 Basel III Accord

Basel III is part of the continuous effort to enhance the banking regulatory framework (BCBS, 2011). It builds on the Basel I and Basel II documents, and seeks to increase the banking sector's ability to deal with financial stress in 2008, improve risk management, and strengthen the banks' transparency. The enhancements of Basel III over Basel II come primarily in four areas augmentation in the level and quality of capital, introduction of liquidity standards, modifications in provisioning norms and better and more comprehensive disclosures (Roy, Bindya, & Swati, 2013). Basel III adds new adequate capital rules to protect banks and improve control of liquidity risk. The accord requires even more risk management for banks and increases bank supervision. There were three main problems exposed during the last credit crisis: high leverage ratio, pro-cyclical consequences, and liquidity risk.

To overcome the problem resulted from leveraged ratio, capital structures were redefined, and capital requirement increase.

Under Basel III, there are two types of capital: Tier 1 equity capital (Common Equity Tier 1 Capital[14] and Additional Tier 1 Capital) and Tier 2 capital. Tier 3 capital in Basel II no longer exists. Tier 1 equity capital must be at least 4.5% of RWA at all times. Total Tier 1 capital must be at 6% of RWA at all times. Total capital must be at least 8% at all times[15].

In addition to new capital requirement, Capital Conservation Buffer (CCB) needs to be considered to protect banks from recessions or financial crises in normal times so that it can be used for making up for the losses incurred during the financial crisis. Banks are required to set up a further 2.5%[16] of RWA to add in Tier 1 equity capital, but this additional requirement can be ignored when in stressed market conditions and once financial markets stabilize, banks will face pressure to increase the ratios again. The idea behind the buffer is that it is easier for banks to raise equity capital in normal periods than in periods of financial stress. The buffer will be phased in between January 1, 2016, and January 1, 2019.

The new regulation will also reduce the effect from pro-cyclicality by considering systemic risk. A common explanation for the pro-cyclicality of the financial system has its roots in information asymmetries between borrowers and lenders (Borio and Lowe, 2001). When economic conditions are depressed and collateral values are low, information asymmetries can mean that even borrowers with profitable projects find it difficult to obtain funding. When economic conditions improve and collateral values rise, these firms are able to gain access to external finance and this adds to the economic stimulus.

---

[14] Also known as Core Tier 1 Capital.

[15] (BCBS, 2011) paragraph 50.

[16] (BCBS, 2011) paragraph 129.

This explanation of economic and financial cycles is often known as the "financial accelerator". The buffer is intended to provide protection for the cyclicality of bank earnings. There will be more control over securitization, and fewer over the counter (OTC) transactions will be permitted. Countercyclical Buffer is introduced, and it is similar to CCB (up to 2.5% of RWA)[17]. The buffer is intended to provide protection for the cyclicality of bank earnings (Hull, 2012). Like the capital conservation buffer, the countercyclical buffer requirements will be phased in between January 1, 2016, and January 1, 2019.

Basel III has also introduced requirements involving two liquidity ratios that are designed to ensure that banks can survive liquidity pressures. The ratios are Liquidity Coverage Ratio (LCR) for short-term liquid (30 days) and Net Stable Funding Ratio (NSFR) for long term liquid (one year).

The capital requirement for Basel Accord III is summarises in Table 2-6. The Basel III accord requires more transparency and more capital in the long-term reserves for liquidity risk. Basel III dramatically increased the amount of equity capital banks were required to keep. It also recognized that many problems for the banks during the crisis were liquidity problems and imposed new liquidity requirements for financial institutions (Hull, 2012).

In conclusion, Basel Accord is still being updated and enhanced. The core of the Basel Accord is related to the amount of capital requirement. For regulators, it is hoped that banks will have enough capital to deal with various risks. For banks, they can provide a loan gaining more profit if the capital requirements are lower.

[17] (BCBS, 2011) paragraph 142.

Table 2-6 Basel III Accord minimum capital requirement

	General capital requirements	Capital Conservation Buffer (+ 2.5%)	Countercyclical Buffer (+ 2.5%)
Tier 1 equity Capital	4.5 % of RWA	7 % of RWA	9.5 % of RWA
Tier 1 Capital	6 % of RWA	8.5 % of RWA	11 % of RWA
Total Capital	8 % of RWA	10.5 % of RWA	13 % of RWA

Sources: (BCBS, 2017)

2.4.4 Basel Accord about SMEs

Growth of SMEs faces three constrains: limited access to financial requirements, trade barriers, and credit risk (Khorasgani and Gupta, 2017). The primary source of external funding for SMEs is debt from bank lending (N. Berger and F. Udell, 1998; Rostamkalaei and Freel, 2015) and hence forecasting loan performance is a persistent problem to banks and financial institutions. This problem is exacerbated in less favourable economic environments, with the effect that credit may be restricted and over-priced (Rostamkalaei and Freel, 2015). There are therefore detrimental effects for SMEs, banks, and the wider economy if credit risk is incorrectly or inadequately measured. E. I. Altman and Sabato (2007) further suggested that more accurate credit scoring models in the market for SME loans have many potential benefits. First of all, if banks are able to improve the accuracy of their credit scoring models, their capital requirements may be lower. Second, if banks are able to reduce their capital requirements in SME lending, this could result in lower interest rates for their SME customers.

The Basel II accord sets up capital requirements that are more sensitive to risk, to which banks respond by charging higher risk premiums on their SMEs debt portfolio (Schindele and Szczesny, 2016). In addition, Dietsch and Petey (2004)

and Jacobson, et al. (2005) showed that bank loan portfolios of SMEs loans are usually riskier than corporate credit. However, lowering of risk premium might be reasonable if banks are required to maintain lower capital requirement on their SMEs credit portfolio, and this requires banks to show that SMEs credit portfolios are not as risky as perceived by the current Basel accord or by banks themselves. Thus, in order to validate this claim, developing a greater understanding of the process that is used in determining the minimum capital requirements for SMEs debt portfolio is urgent.

The credit risks generated by retail banking are significant, but they have a very different dynamic from the credit risk of commercial and investment banking businesses (Crouhy, Galai, & Mark, 2014). As mentioned earlier, the calculation of capital requirement about credit risk under Basel II Accord depends on the treatment of SMEs (corporates or retails). Basel regulations allow banks to classify their SMEs credit lending as retail (only 75% risk weight under standardized approach) if their overall amount of credit exposure does not exceed one million Euros and stays below 0.2% of their total retail portfolio. This lower the capital requirement for the banks.

In addition, banks are permitted to distinguish separately exposures to SMEs under IRB approach of Basel II and Basel III (defined as corporate exposures where the reported annual sales for the consolidated group of which the firm is a part is less than €50 million) from those to large firms. A firm-size adjustment (i.e.$0.04(1-\frac{S-5}{45})$) is made to the corporate risk weight formula for exposures to SME borrowers. S is expressed as total annual sales in millions of euros with values of S falling in the range of equal to or less than €50 million or greater than or equal to €5 million. Reported sales of less than €5 million will be treated as if they were equivalent to €5 million for the purposes of the firm-size adjustment for SME borrowers[18].

[18] (BCBS, 2006) paragraph 273.

$$\text{Correlation(R)} = 0.12\frac{1-e^{(-50\times PD)}}{1-e^{-50}} + 0.24\left(1-\frac{1-e^{(-50\times PD)}}{1-e^{-50}}\right) - 0.04(1-\frac{S-5}{45}) \quad (16)$$

The risk assessment modifications in the Basel Accord affect the retail and corporate banking sectors differently. For given PD, capital requirement estimated using the retail formula are lower than those estimated using the corporate formula. In turn, lower capital requirement may lead to enhanced supply and lower cost of credit to SMEs (Edward I Altman and Sabato, 2005). Lower capital requirement for retail clients may be due to the fact that, unlike large firms, SMEs are less adversely affected by systemic risks (Dietsch and Petey, 2004). Dietsch and Petey (2004) studied French and German SMEs large datasets, and concluded that SMEs carry higher risk, and the asset correlations in the SMEs population are very weak (1–3% on average) and decrease with size. This means banks will have significant benefits, in terms of lower capital requirements, when considering SMEs as retail. But they will be obliged to use the A-IRB approach and to manage them on a pooled basis.

In the case of SMEs as retails, banks using A-IRB approach must provide their own estimates of PD, EAD, and LGD. This would be a challenge of the ability of a bank's internal risk rating system to adequately capture the differences between different loans and different types of assets, and the methods used to calculate the relevant risk measures (Jacobson, et al., 2005). Edward I Altman and Sabato (2005) found that majority banks will use a blend approach (considering some SMEs as retail and some as corporate). Through a breakeven analysis, they found that banking organizations for US, Italy and Australia will be obliged to classify as retail at least 20% of their SME portfolio in order to maintain the current capital requirement (8%). Banks will face the choice of either increasing regulatory capital requirements for SMEs' risk exposure or increasing organizational and technological costs to meet the Basel II risk management standards for SMEs' retail treatment.

Hence, the estimation of the PD becomes a key aspect of the new banking regulation using IRB approach. It must be a long-run average of 1-year default rates for borrowers in the grade. The length of the underlying historical observation period used must be at least 5 years, and the bank is permitted to apply for its calculation by one or more of the following techniques: i) internal default experience; ii) mapping to external data; or iii) statistical default models[19].

Additionally, on December 7, 2017, BCBS published a document finalizing the Basel III reforms, also known informally as Basel IV (BCBS, 2017). Basel IV applied some adjustment to enhance the accuracy of capital requirement estimation for SMEs. Under the standardized approach, SMEs are separately identified. For unrated exposures to corporate SMEs (defined as corporate exposures where the reported annual sales for the consolidated group of which the corporate counterparty is a part is less than or equal to €50 million for the most recent financial year), an 85% risk weight will be applied. Exposures to SMEs that meet the criteria of regulatory retail SME exposures and risk weighted at 75%[20]. In terms of the risk components, the floor of PD increases to 0.05% under IRB approach[21]. Yet, SMEs have not been separated from large companies. The foundation of the model is based on a sample of large firms, which may not lead to accurate results for SMEs. For instance, the LGD used in F-IRB estimation is still equal for both large firms and the SMEs sample.

Basel III will be fully implemented in 2019. Under the new banking regulation, the way an SME is treated will differ according to the approach chosen by the particular bank, Standardised or IRB, and according to whether the bank includes the SME in the corporate or retail category. However, model risk may

[19] (BCBS, 2006) paragraph 461.

[20] (BCBS, 2017) Standardised approach for credit risk paragraph 43, 55.

[21] (BCBS, 2017) Internal ratings-based approach for credit risk, paragraph 68.

cause increased correlations in bank returns, engendering cyclical fluctuations in the financial condition of the banking sector, with potentially macroeconomic consequences (Allen, DeLong, & Saunders, 2004).

2.5 International Financial Reporting Standard (IFRS) 9

Credit risk has been identified as a major cause of the credit crisis starting in 2007. The delayed recognition of loan losses on part of banks and other lenders under the incurred loss model approach of International Accounting Standard (IAS) 39 has been heavily criticised as a major weakness of financial accounting standards (Huian, 2012) since the beginning of the financial crisis. People realize that the warning of the loan loss that should have been provided earlier was far insufficient ("too little, too late"). Furthermore, several high-profile groups have argued that the incurred loss approach reinforces the pro-cyclical effects of bank regulation. The pro-cyclical mechanism of the incurred loss model leads to rather lower impairments, and therefore higher gains, during economic booms. In downturns, however, it initially causes small write-downs, though, these are followed by massive ones (Novotny-Farkas, 2016).

In April 2009, the Financial Stability Forum (Financial Stability Board, 2009) and the G-20 state leaders (Summit, 2009) sent urgent requests to the International Accounting Standards Board (IASB) and the Financial Accounting Standards Board (FASB) to improve their impairment rules. The G-20 leaders, investors, regulatory bodies and prudential authorities required standard setters to develop a new accounting standard that allowed for a more forward-looking provisioning, and to propose the improvement of the standard for financial instruments with the view to increase financial stability, taking into account:

- the complexity of the existing standard for financial instruments,
- the extent to which the financial instrument is subject to fair value,
- the procedure of recognition and measurement of financial instruments.

In response, the International Accounting Standards Board (IASB) has devoted considerable effort to resolving this issues, and published the final version of IFRS 9 Financial Instruments in July 2014 (IASB, 2014). The Standard includes requirements for recognition and measurement, impairment, de-recognition and general hedge accounting. The new impairment model in IFRS 9 has come into effect in replacement of the incurred loss impairment model of IAS 39 (IASB, 2003). The IASB's Chairman, in a speech in January 2016 before the European Parliament, pointed out that the biggest change deriving from the replacement of the standard is a model of expected credit loss that requires a timely recognition of inevitable losses in financial statements, particularly in banks (Gornjak, 2017). With the prospective impairment criterion of expected credit loss, any expected defaults in the future, taking into account all relevant internal and external information, should be anticipated and reassessed at each reporting date.

The IFRS 9 is a principle-based and logical rather than rule-based, and its objective is to establish financial reporting principles on financial assets and liabilities to present important and useful information to the users of financial statements in order to estimate the amount, time, and uncertainty of the entity's future cash flows (Tominac and Vašiček, 2018). A new standard for banks can contribute to improving credit risk management, increasing transparency regarding asset quality and credit risk, and reducing pro-cyclicality through more timely recognition of credit losses (Frykström and Jieying, 2018).

2.5.1 From Incurred Loss (IAS 39) to Expected Credit Loss (IFRS 9)

Credit loss is defined as the present value of differences between all contractual cash flows and the cash flows expected to flow in ('cash shortfalls') discounted at the initial effective interest rate. It should be noted that a credit

loss already occurs when contractual payments expected to arrive are delayed[22].

As shown in Figure 2-3, fair value accounting (FVA) expected loss approach is the most comprehensive since it accounts for all risk factors. Fair value accounting uses current market values as the basis for recognizing certain assets and liabilities. Fair value is the estimated price at which an asset can be sold, or a liability settled in an orderly transaction to a third party under current market conditions and assets and liabilities are re-measured periodically to reflect changes in their value.

Incurred losses are expected losses from events as the balance sheet date. Thus, incurred losses represent a subset of expected losses. Expected credit losses are incurred credit losses and expected credit losses from events expected to occur after the balance sheet date (G. Gebhardt, 2016). The incurred loss model requires the recording of credit losses that have been incurred as the balance sheet date, rather than probable future losses, and the loss identification is based on the occurrence of triggering events supported by observable evidence (e.g. borrower loss of employment, decrease in collateral values, past-due status) combined with expert judgment (B. H. Cohen and Edwards, 2017). The IFRS 9 expected loss model is positioned between the IAS 39 incurred loss approach and FVA, because it recognises ECL but ignores changes in market interest rates (Novotny-Farkas, 2016).

Most entities monitor the financial assets and recognise an impairment only when there is objective evidence of default or a particular balance is past due beyond a certain point. An entity only considers those impairments that arise as a result of incurred loss events. It is not permitted to reporting entities to subjectively consider expected losses. The explanation is that prudent

22 (IASB, 2014) Pages A414, Paragraph B5.5.28.

recognition of loan losses could have potentially decreased the cyclical moves in the financial crisis (Bushman and Landsman, 2010).

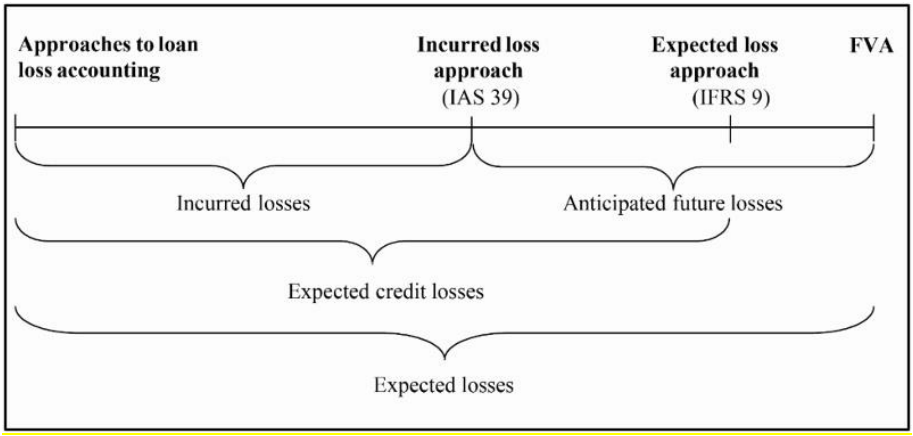


Figure 2-3 Loan loss recognition under alternative accounting regimes taken from G. u. Gebhardt and Novotny-Farkas (2011)

Under IAS 39, the impairment model further excludes expected credit losses from events expected to occur after the balance sheet date. For example, if a major employer in a region announces to close a factory in the next fiscal year and to lay off thousands of employees, this will result in additional credit losses for local banks. These additional expected credit losses may not be recognised as the event (i.e. the closure of the factory) does not take place before the balance sheet date but only in the next fiscal period (G. Gebhardt, 2016). Loss allowances were only recorded for impaired exposures, and therefore, this leads to the delay of recognition of loss.

With regard to the IFRS 9 expected credit loss model, objective evidence of the existence of a loss event is not required for recognising impairment at initial. It is based on the rationale that initially expected credit losses over the maturity of the debt instrument are reflected in a credit risk premium included in the interest rate. If expectations about credit losses increase, this should be covered by setting up a loan loss allowance for lifetime expected credit losses.

The new IFRS 9 impairment requirements eliminate the IAS 39 threshold for the recognition of credit losses, i.e., it is no longer necessary for a credit event

to have occurred before credit losses are recognised (Figure 2-4). This is a shift from accounting based to risk based models (Ozdemir, 2018). Instead, an entity always accounts for expected credit loss (ECL) and updates the loss allowance for changes in these ECL at each reporting date to reflect changes in credit risk since initial recognition. Consequently, the holder of the financial asset needs to take into account more timely and forward-looking information. The new impairment requirements result in earlier recognition of credit losses, by necessitating a 12-month ECL allowance for all credit exposures not measured at fair value through profit or loss. In addition, there will be a larger allowance for all credit exposures that have significantly deteriorated (as compared to the recognition of incurred losses under IAS 39 today). While credit exposures in stage 3, are similar to those deemed by IAS 39 to have suffered individual incurred losses, credit exposure in stages 1 and 2 will essentially replace those exposure measured under IAS 39's collective approach.

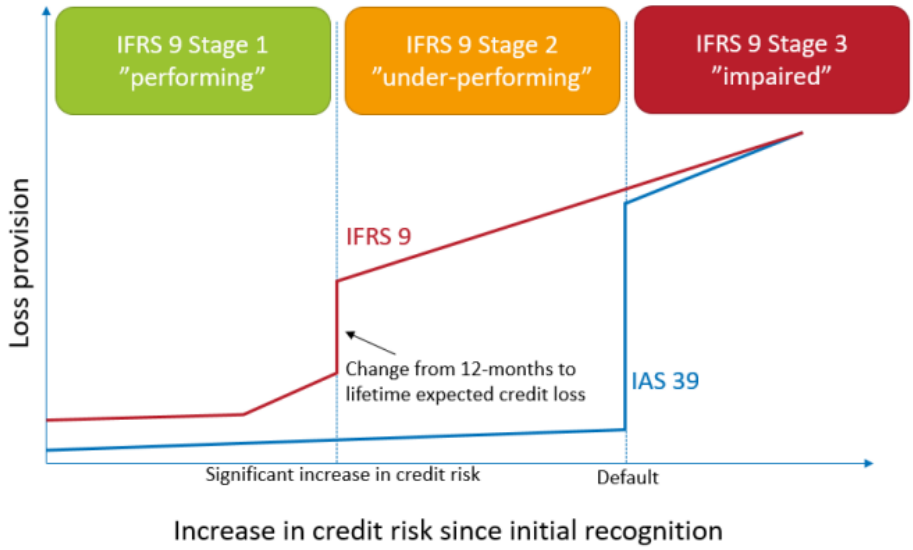


Figure 2-4 Development of provisions under IFRS 9 and IAS 39 taken from (Frykström and Jieying, 2018)

While expected credit loss model broadens the information. An entity is required to consider all relevant information which reflects the effects of current conditions including historic, current and forward-looking information even macroeconomic data. Besides, the entity should also consider reasonable and

supportable information that includes forecasts of future economic conditions including, where relevant, multiple scenarios, when calculating ECL, on an individual and collective basis. This will result in the earlier recognition of credit losses as it will no longer be appropriate for entities to wait for an incurred loss event to have occurred before credit losses are recognised (G. Gebhardt, 2016). Also, IFRS 9 requires significant enhancements to a financial institution's data, systems, quantitative models and governance.

2.5.2 Impairment Model Under IFRS 9

Every loan and receivable have some risk of defaulting in the future, every loan or receivable has an expected credit loss associated with it from the moment of its origination or acquisition. Impairment model is the biggest change for financial institutions moving from IAS 39 to IFRS 9. With the IFRS 9 standards, impairment recognition in loans and receivables that are measured at Amortized Cost or Fair Value through Other Comprehensive Income (FVOCI) will follow a forward-looking expected credit loss (ECL) model (IASB, 2014).

Under the ECL approach, at each reporting date, an entity recognises a loan loss allowance (also known as loan loss provision) based on either 12-month ECL or lifetime ECL, depending on whether there has been a significant increase in credit risk on the financial instrument since initial recognition[23].

Credit exposures will be categorized into one of three stages, depending on the levels of credit risk since initial recognition or at reporting date. An entity must determine whether the financial asset is in one of the three stages in order to determine both the amount of ECL to recognise as well as how interest income should be calculated. In Table 2-7, the three-stage general approach in the IFRS 9 impairment model reflect the general pattern of the deterioration in credit risk of a financial instrument that ultimately defaults. At each reporting

[23] (IASB, 2014) Pages A321

period, an entity assesses which stage a financial instrument that is subject to impairment testing falls into. The stage determines the relevant impairment requirements. According to the change in credit risk, the financial instrument is placed on stage 1 or stage 2 or stage 3. If the credit risk for a financial instrument improves, the instrument can move from Stage 2 back to Stage 1 but movements from Stage 3 back into Stage 2 or Stage 1 are rarer and harder to justify.

Stage 1 includes financial instruments with no significant increase in credit risk since initial recognition, or financial instruments that have low credit risk at the reporting date. For these assets, the 12-month ECL is calculated and recognized as a provision in liability in the statement of financial position and as profit or loss in the income statement. Interest income must be recognised on the basis of the gross carrying amount. This means that interest will be calculated on the gross carrying amount of the financial asset before adjusting for ECL. Unless its credit quality changes, the same treatment will then apply every reporting date until its maturity

Stage 2 includes financial instruments with a significant increase in the default risk, but no objective evidence of impairment. IFRS 9 requires that when there is a significant increase in credit risk, institutions must move an instrument from a 12-month expected loss to a lifetime expected loss. In making the evaluation, the institution will compare the initial credit risk of a financial instrument with its current credit risk, taking into consideration its remaining life. This is because an economic loss arises when ECL significantly exceed initial expectations. The resulting increase in the provisions is typically expected to be significant. As in stage 1, interest income will continue to be recognised on the basis of the gross carrying amount according to the effective interest method.

.

Table 2-7 Overview of the general IFRS 9 impairment approach

Loss allowance updated at each reporting date	Stage		
	Stage 1 →	Stage 2 →	Stage 3
Level of credit risk deterioration	No significant increase in credit risk since initial recognition	Significant increase in credit risk since initial recognition	Credit-impaired
Impairment recognition	12-month ECL	Life-time ECL	
Interest rate to apply	Effective interest rate on gross carrying amount before allowance for ECL		Effective interest rate on gross carrying amount after allowance for ECL

The effective interest rate is the rate that discounts the estimated future cash flows from the asset to the asset's Amortized Cost before any allowance for expected credit losses.

Stage 3 includes financial assets which are considered to be credit-impaired when one or more events that have an unfavourable impact on its estimated future cash flows have occurred at the reporting date. If objective evidence of impairment exists, this is effectively the point at which there has been an incurred loss event under the IAS 39 model. Entities will continue to recognise lifetime ECL but they will now recognise interest income on a net basis. This means that interest income will be calculated based on the gross carrying amount of the financial asset less ECL. In subsequent reporting periods, if the credit quality of the financial asset improves so that the financial asset is no longer credit-impaired and the improvement can be related objectively to the occurrence of an event (such as an improvement in the borrower's credit rating), then the entity should once again calculate the interest revenue by

applying the effective interest rate on the gross carrying amount of the financial asset. The recognition of lifetime ECL will occur earlier than under IAS 39, that is, already when there is a significant increase in credit risk (Stage 2), but before actual default (Stage 3).

Besides, there is a simplified approach for qualifying trade receivables, contract assets within the scope of IFRS 15 and lease receivables. These assets must recognize a loss allowance based on lifetime ECL and effective interest rate on gross carrying amount (or gross carrying amount after allowance for ECL if it is credit-impaired) at each reporting date rather than the general approach.

The credit adjusted approach applies only rarely when an entity acquires or originates a loan or receivable that is “credit impaired” at the date of its initial recognition (e.g., when a loan is acquired at a deep discount due to credit concerns via a business combination). An asset is credit impaired when one or more events that have a detrimental effect on the estimated future cash flows of the asset have occurred. This asset must suffer a loss allowance based on credit adjusted effective interest rate, which differs from the effective interest rate in that estimates of future cash flows includes an adjustment for expected credit losses.

2.5.3 Significant Increase in Credit Risk(SICR)

The main trigger within the general model falling into stage 2 is a significant increase in credit risk. IFRS 9 does not quantitatively define significant increase in credit risk, but there are multiple criteria to determine if an account has to be moved to Stage 2.

IFRS 9 suggests that investment grade rating might be an indicator for a low credit risk. This is to say judgement of credit risk’s significant change can be done by the change in internal or external rating, such as Fitch, Moody and Standard & Poor's. Second is to compare to credit risk at initial recognition as

a relative rather than an absolute assessment. The assessment of whether there has been a significant increase in credit risk is based on an increase in the probability of default occurring since initial recognition. An entity is allowed to use various approaches to assess whether credit risk (probability of default) has increased significantly at the reporting date provide that the approach is consistent with the requirements. Third is about the definition of default and maturity. The requirement of IFRS 9 contain a rebuttable presumption that credit risk has increased significantly when contractual payments are more than 30 days past due. This means if missing one-month payment, an account is likely to be moved into Stage 2 where additional interest or collateral is required based on the existing contract with the borrowers. Yet, banks can reject this move, and they basically have to build a model to argue that even though the borrower is one-month past due, but because of behavioural patterns, it may be acceptable if a little bit more than 30 days. In general, banks are likely to rebut retail loans rather than corporate loans. If corporates miss a payment, it is usually not a good sign and an indication that they are underperforming. It would be difficult to rebut corporate and SMEs accounts (Raj, 2016). An entity cannot rely solely on past due information[24]. Fourth is that adverse macroeconomic changes will not have an impact on the borrower's capacity to honour its obligations. Finally, reasonable and supportable information from internal and external of an entity should be considered.

Deloitte carried out a series of questionnaires to study how financial institutions define the significant change of credit risk in practice. 41% of the bank questioned defined as a trigger the missed payments and 35% the change in in the rating (Deloitte., 2015). Besides, Novotny-Farkas (2016) pointed out that a significant proportion of financial assets that are currently disclosed under the label 'Financial assets past due, but not impaired' in bank financial

[24] (IASB, 2014) Pages A322, Paragraph 5.5.11.

statements would largely fall into Stage 2 under IFRS 9. PwC[25] suggested that an entity has to build up an accounting policy to judge whether an increase in credit risk in the context of its own internal risk management and reporting.

2.5.4 12-month ECL and Lifetime ECL

Depending on the stage that the instrument falls into, an entity either recognizes 12-month ECL (stage 1) or lifetime ECL (stage 2 or stage 3). As ECL consider both the amount and the timing of payments, a credit loss arises even if the holder expects to receive all the contractual payments due, but at a later date. For example, banks now have to take an impairment loss because they reluctant to make impairment provision on customers who, though, paid in full but much later than the due date. This is because the timing of payments directly affects present value and thus the amount of impairment loss under IFRS 9 (Deloitte, 2017).

Besides, ECL are also a probability-weighted estimate of credit losses over the expected life of the financial instrument, and the standard further defines ECL as the weighted average of credit losses with the respective risks of a default occurring as the weights. Every receivable account carries with it some probability of default and therefore has an ECL attached to it. When measuring ECL, an entity needs to consider[26]:

- An unbiased and probability-weighted amount that is determined by evaluating a range of possible outcomes,
- The time value of money,
- Reasonable and supportable information about past events, current conditions and forecasts of future events and economic conditions at the reporting date.

---

[25] (PWC, IFRS 9, Financial Instruments Understanding the basics, Pages 32, https://www.pwc.com/gx/en/audit-services/ifrs/publications/ifrs-9/ifrs-9-understanding-the-basics.pdf)

26 (IASB, 2014) Pages A324, Paragraph 5.5.17.

ECL are updated at each reporting date for new information and changes in expectations. The ECL estimate should reflect an impartial and probability-weighted amount that is determined by evaluating a range of possible outcomes.

A straightforward and commonly applied approach to valuing the ECL under IFRS 9 is a probability-weighted loss default (PLD) model. The PLD model involves the following four key parameters. Probability of default (PD), which is the likelihood of a counter-party default (failure to meet repayment/debt obligations) during a particular period of time. Exposure at default (EAD), which is the total value that one entity is exposed to when a counter-party defaults or outstanding and unsecured credit amount at the event of default. Loss given default (LGD), which is the percentage of contractual claims that would be lost if the counter-party defaults, and usually is defined as a percentage of the exposure at default. Discount factor (DF), which is the factor that needs to be multiplied in order to convert future cash flows into the present value at the measurement date. In this model, the ECL is derived by summing the ECL of all the expected default events within a specific period (either 12 months or a lifetime). The ECL for each possible event is calculated as the product of the four parameters above, through the formula shown below:

$$\mathrm{ECL} = \sum_{i=1}^{T} \mathrm{PD}(t_i) \times \mathrm{LGD}(t_i) \times \mathrm{EAD}(t_i) \times \mathrm{DF}(t_i) \qquad (17)$$

where t_i refers to a time factor.

The computation of credit losses over a lifetime horizon is one of the key innovations introduced by the new accounting standards (Bellini, 2019). IFRS 9 defines lifetime ECL as the ECL that result from all possible default events over the expected life of a financial instrument (i.e. an entity needs to estimate the risk of a default occurring on the financial instrument during its expected life). In other words, lifetime ECL is the expected present value of losses that arise if borrowers default on their obligations at some time during the life of the

financial asset, while 12-month ECL is referred to the portion of the lifetime ECL that result from possible default events within 12 month after reporting date, rather than the expected cash shortfalls over the next 12 months. 12-month ECL should be recognized even if there has not been a significant increase in credit risk, while lifetime ECL should be recognized when there is a significant increase in credit risk or evidence that credit impaired has been determined. It is reasonable to treat 12-month ECL and lifetime ECL separately even though one-year ECL can be seen as a snapshot of lifetime because banks have been developing one-year models for a relatively long period due to the Basel Accord regulatory requirement (Bellini, 2019).

However, IFRS 9 neither specifies nor recommends any specific methodology to measure the 12-month or lifetime ECL. It does not provide any detailed steps for deriving the parameters of any one selected approach. In order to calculate 12-month and lifetime ECL, banks should apply models on credit risk (PD, LGD), balance sheet forecast (prepayments, facility withdraws) and interest rates (discount factors). On the credit risk side, PD and LGD models are needed to satisfy the new impairment model. Besides, although the definition in IFRS 9 of 12-month ECL is similar to the Basel Committee's definition of ECL, the modelling requirements differ significantly because the IFRS 9 measure is a point-in-time estimate, reflecting currently forecast economic conditions, while the Basel regulatory figure is based on through-the-cycle assumptions of default and conservative estimates of losses given default.

Finally， the impairment model aims to achieve an appropriate balance between faithful representation of ECL and the operational costs and complexity. The 12-month ECL is a simplification included in IFRS 9 due to the cost-benefit of the lifetime ECL requirement. Additionally, 12-month ECL are already being computed by some regulated financial institutions, hence, implementing this requirement would be less costly (Ernst&Young, 2018).

2.5.5 The Impact of IFRS 9 on Banks and SMEs

Both financial institution and non-financial institution need provision to cover eventual asset impairment and potential obligations that have still not materialized. For financial institution such as banks, default provision has an important effect on the operation. Every time a bank makes a loan, it has to set up the corresponding provision to cover the risk that a borrower may default or fall behand in their payment obligation, which aims to reduce the credit risk (Tominac and Vašiček, 2018). Intuitively, it makes sense that a high-risk weight means a large portion of provision requirement increases as impairment rises and if the credit quality of financial assets improves the recognized provision is reverted.

Since the banks are the biggest loan providers the implementation of ECL in IFRS 9 would have far-reaching implication to globally financial industry, and significantly affects both banks and borrowers (Tominac and Vašiček, 2018). It is expected that impairment provisions could be 20-250 percent higher under IFRS 9 compared to IAS 39 (Deloitte, 2017), and the biggest impact would be felt during the transition period from IAS 39 to IFRS 9. The substantially increase in provisions negatively affects profit and loss statements (earnings) and weight on the capital, thus affecting regulatory capital (Deloitte, 2016b) and it would also potentially affect dividend pay-outs as well. Therefore, it is possible to obtain a loan at a higher cost depending on credit history of borrowers because of the forward-looking requirement from IFRS 9. Bank will have to recognise provision from the day they extend any loan in anticipation of future potential losses instead of the present situation only. In order to deal with the larger amount of provision, banks have to revaluate or restructure the loans by building a robust risk profile with more complete and useful information from potential debtors, making it more expensive to borrowers, especially with those who is high risk. In view of this, treatment of missing data becomes important and SMEs with lots of missing data would negatively impact on their risk profile building.

There is again with the three-stage model for impairment according to significant change in credit quality since the day the loan was extended. When banks do provision, they have impaired and non-impaired accounts under IAS 39, while the non-impaired accounts have to be split into one of two stages: Stage 1 (performing account) and Stage 2 (underperforming account). Stage 3 (nonperforming account) is for impaired loans. As mentioned earlier, for accounts that fall under Stage l, borrowers with good credit risk profile will likely fall under Stage 1, and the bank has to provide a 12-month forward-looking ECL. The provision heavily increases if the account gets into Stage 2, which is potentially around four to five times more than that for Stage 1, and lifetime ECL has to be provided. For example, if a retail mortgage loan has an expected maturity of 20 years and it has gone into Stage 2, the bank has to provide (over) 20 years ECL, instead of 12 months. Different loan term carries different provisioning. The longer tenure provisioning will be higher. This is also the main difference between the provisions recognized under the IFRS 9 and IAS 39. In the past, banks would give a loan and like to stretch it to a longer loan term because it gives them recurring income, now they will have to carefully consider borrowers with poor rating or credit history.

In view of the above, in the face of interest rate caps and prudent reporting required by IFRS 9, SMEs may be caught in the middle. SMEs to get a longer tenure loan can be more expensive leading a significant impact on their survival. It is challenging for SMEs to lower the risk, and fundamentally change the way of calculating bad debt provision for receivables from an incurred loss to an expected loss model and make a provision charge from day one.

Finally, IASB proposed an IFRS for SMEs to ease the financial reporting burden as well as reduce the cost of financial information disclosure. The IFRS for SMEs Standard is tailored for small firms. It is beneficial for SMEs to provide information such as cash flow, liquidity and solvency to lenders, creditors and

so on.[27]. There is roughly a 90% reduction in disclosures in comparison with full IFRSs.

All in all, the IFRS 9 is forward-looking and ensures a more accurate, and timely assessment of expected losses, and it has great impacts on banks, but it also influences on SMEs with debt and or receivables account for determining allowance.

2.6 Basel Accord and IFRS 9

Accounting reflects the nature of the financial asset (determined by its cash flow characteristics), the company's business model (how the assets are managed) and its risk management practice on financial statements.

Since regulators use financial statement information to calculate regulatory capital numbers and rely on market participants to trade on this information to discipline banks, financial reporting and bank supervision are closely intertwined. Supervisors' primary objective is to reduce the level of risk to which depositors are exposed and to maintain financial stability. In contrast, the main objective of financial reporting is to provide information that is useful to a wide range of financial statement users, including investors, creditors and regulators.

2.6.1 Rating Philosophies

Under both Basel III and IFRS 9 frameworks, the key input parameters for the measurement of expected loss are the PD. Yet, PD for regulatory purpose cannot be applied directly to expected credit losses impairment calculations under the IFRS 9 (Conze, 2015).

[27] The IFRS for SMEs Standard, https://www.ifrs.org/issued-standards/ifrs-for-smes/

The rating philosophy for PD modelling may follow a point-in-time (PIT), through-the-cycle (TTC) or a hybrid approach. The PIT PD assesses the likelihood of default over a relatively short horizon. As it assesses risk at a point in time, the borrower will move up or down rating grades through the economic cycle. It can fluctuate dramatically over the business cycle. In contrast, TTC PD predicts average default rate performance for the borrower over an economic cycle and ignores short run changes to a customer's PD. Its estimates therefore lead to a more smoothed and less cyclical ratings. The hybrid approach is a combination of TTC and PIT models, which means that PD ratings are calibrated to long run default rates but adjusted to reflect current economic conditions (Mayer, Resch, & Sauer, 2017).

The nature of the PD model is usually determined by the degree of cyclicality in the underlying model drivers. Figure 2-5 illustrates how PD estimates vary over the business cycle depending on the underlying rating philosophy. TTC estimates are likely to overstate PIT PD during boom periods (PD going down) and understate PIT PD during downturns (PD going up). A PIT rating system vary more significantly in expansionary to recessionary periods and is generally prevalent in day-to-day risk management of retail portfolios. PIT and TTC PD are extremely stylised and in reality, many rating systems are hybrid approaches.

The Best practice of estimating PD under Basel III must consider the following elements[28]:

- A long-run average estimates of PD (one-year PD)
- The data should include a representative mix of good and bad years of the economic cycle relevant for the portfolio.
- PD must be forward-looking – a simple extrapolation from historical data is only a starting point

[28] (BCBS,2017) Standardised approach for credit risk paragraph 230, 231.

These requirements would suggest that only TTC or hybrid approaches are consistent with the capital adequacy framework.

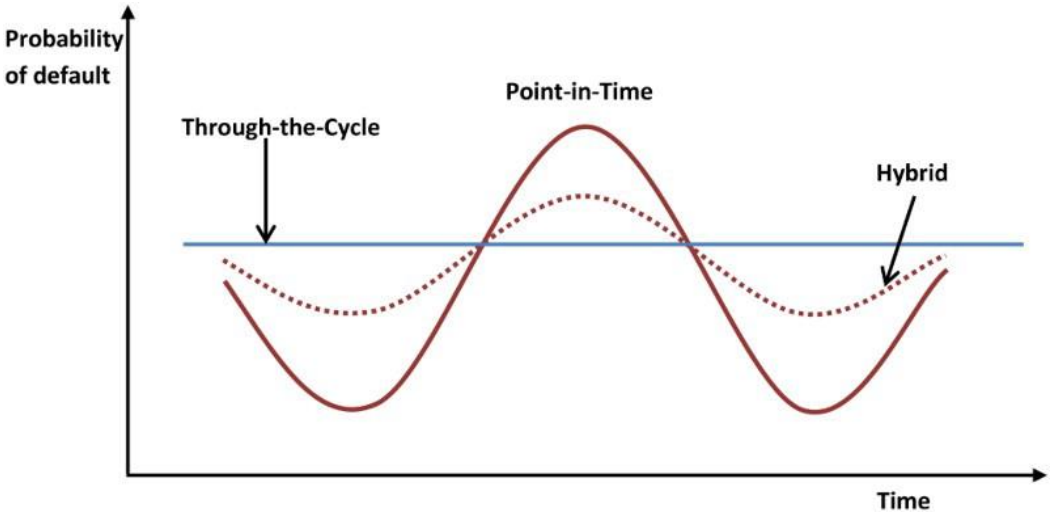


Figure 2-5 PD of TTC versus PIT over the business cycle taken from Doorsselaere (2015)

The IASB clarifies that TTC estimates are not consistent with IFRS 9 expected loss requirements because IFRS 9 standards require a PD estimate that is consistent with the following principles:

- Considers all relevant information
- Reflects current economic circumstances (i.e., it is a best estimate rather than a conservative estimate)
- Provides the likelihood of a default occurring within the next 12 months or during the lifetime of the instrument
- Includes forward-looking economic forecasts
- Existing internal ratings-based (IRB) Basel models can be reused but particular attention should be paid to point-in-time versus through-the-cycle models

IFRS 9 consider a range of possible economic outcomes instead of those that are actually expected at the reporting date[29]. TTC would result in a loss

[29] (IASB, 2014) Pages A324, Paragraph 5.5.17.

allowance that does not reflect the economic characteristics of the financial instruments at the reporting date. Therefore, PD estimates based on PIT measures of current and expected future conditions reflecting future economic cycles at the reporting date.

Under Basel Accord, PD measures the average of default within the next 12 months while Under IFRS 9, depending on the asset, the PD measures either for the next 12 months (stage 1) or for the remaining life of the financial instrument (stages 2 and 3). For defaulted assets, lifetime losses have to be recognised under both frameworks.

Most banks subject to IFRS 9 are also subject to Basel III Accord capital requirements and, to calculate credit risk-weighted assets, use either standardized or internal ratings-based approaches. The new IFRS 9 provisions will affect the Profit and Loss of the financial institution that in turn needs to be reflected in the calculation for impairment provisions for regulatory capital. The infrastructure to calculate and report on expected loss drivers of capital adequacy is already in place. The data, models, and processes used today in the Basel framework can in some instances be used for IFRS 9 provision modeling, albeit with significant adjustments. Banks that use an advanced approach to calculate their capital requirements should be able to use their existing systems and methodologies as a starting point and make the necessary adjustments to flex the calculation to comply with IFRS 9. Not surprisingly, a Moody's Analytics survey conducted with 28 banks found that more than 40% of respondents planned to integrate IFRS 9 requirements into their Basel infrastructure (Temim, 2016).

In summary, it is not surprising that the IASB expects entities to be able to use some regulatory measures as a basis for the calculation of expected credit losses in accordance with the requirements in IFRS 9. However, because of the different objectives of regulation and financial reporting, the regulatory estimates of PD are not the same as those that shall be used for expected loss

calculation of expected losses under IFRS 9. Hence, these estimates have to be adjusted to meet the measurement requirements of IFRS 9.

2.6.2 Capital Ratio and Provisions

Financial losses for banks are uncertain ahead of time. Borio, Furfine, & Lowe (2001) found that lending rates for a loan did not accurately reflected credit risk so that banks will set aside a specific amount as a cushion to absorb expected loss on banks' loan portfolio and this amount is referred to as loan loss provisions (LLPs), which is a credit risk management tool used by banks to mitigate expected losses on bank loan portfolio (Ozili and Outa, 2017). In addition to provisions for the expected losses, banks also hold capital in case losses are larger than expected (unexpected losses) (B. H. Cohen and Edwards, 2017). Yet, the distinction between provision and capital is still confused but it is believed that the sum of them should be able to cover the expected and unexpected losses.

The measurement of loan loss provisions is directly linked to capital ratio calculations. IFRS 9 requires an institution to immediately recognize a 12-month ECL from a financial asset at the first reporting date after origination and create an allowance to cover such loss. The expected credit loss is to be covered by provisions, and unexpected loss is to be covered by capital. Consequently, loss provisions will significantly increase under IFRS 9, thus reducing the equity and retained earnings available for Tier 1 capital, which in turn may reduce the Tier 1 capital ratio.

2.7 Definition of Default

The notion of default is fundamental to the application of the model, particularly because it affects the subset of the population that is subject to the 12-month ECL measure. Basle Committee on Banking Supervision (BCBS, 2006) defined default as any credit loss event associated with any obligation of the obligor, including distressed restructuring involving the forgiveness or

postponement of principal, interest, or fees and delay in payment of the obligor of more than 90 days. According to the current proposal for the new capital accord banks will have to use this tight definition of default for estimating internal rating-based (IRB) models.

For IFRS 9, default is not defined for the purposes of determining the risk of a default occurring, because it is defined differently by different institutions (for instance, 30, 90 or 180 days past due), the IASB was concerned that defining 'default' could result in a definition that is inconsistent with that applied internally for credit risk management. Since the default is the anchor point used to measure probabilities of default and losses given default in Basel modelling, requiring a different definition would require building a different set of models for accounting purposes. It should be mentioned that analysis for different definitions of "default" will produce a different sample count for "default" accounts. Therefore, the standard requires an entity to apply a definition of default that is consistent with how it is defined for normal credit risk management practices, consistently from one period to another. Furthermore, an entity might have to use different default definitions for different types of financial instruments.

The standard restricts diversity resulting from this effect by establishing a rebuttable presumption that default does not occur later than when a financial asset is 90 days past due. This presumption may be rebutted only if an entity has reasonable and supportable information to support an alternative default criterion. A 90-day default definition would also be consistent with that used by banks for the advanced Basel II regulatory capital calculations with a few exceptions. A report from Ernst&Young (2018) suggested that most banks intend to align their regulatory and accounting definitions of default. This generally means aligning the number of days past due trigger to 90 days under IFRS 9, with some exceptions for certain portfolios such as mortgages for which the regulatory definition may allow longer delinquency periods.

2.8 Summary

In this chapter, the author has reviewed the works of literature about missing data, credit risk models, credit risk in Basel Accord and impairment model in IFRS 9. First, there is a major influence on accuracy of statistical analysis if missing data exist. Yet missing data is inevitable in the field of SMEs study, as their data are not as transparent as listed entities. The idea behind how to solve the missing observations problem becomes an essential procedure. Before studying missing data methods, it is necessary to consider the nature of missing data. This depends on the underlying missing data mechanism with respective data as it explains the reason why the data is not observed. Model-based methods (e.g., ML, MI), depended on MAR-data, could be an appropriate approach to solve missing data problem in relation to SMEs data. To summarize, both ML and MI are advanced statistical methods to handle missing data, but in the field of credit scoring, MI is a more reasonable and flexible option. Since binning method has been popular in credit scoring but it is difficult to interpret when dealing with non-linearity, MICE as an alternative to impute the missing data without transformation and to my knowledge, there is little study comparing MICE with the binning methods.

Traditional credit scoring models based on fundamental analysis to find which factors are important in assessing the credit risk of a firm. They evaluate the significance of the pre-identified factors, mapping a reduced set of accounting variables, financial ratios and other information into a quantitative score. As a classical statistical method, logistic regression is unsurprisingly used in this SMEs studies so as to help reduce bank reserves, enhance credit risk management and finally help SMEs boost. Since there is a significant impact on modelling credit risk from the new regulations, the economic cycle, multiple periods and other relevant factors into consideration need to be considered when establishing modelling.

New Basel accord (Basel III or Basel IV) will be fully implemented in the near future. Tightening capital requirement indeed pose a challenge to SMEs'

survival, due to the IRB-approach, it is feasible to calculate the default probability to evaluate the risk to grant credit to SMEs. In terms of IFRS 9, the three-stage model provides a new aspect to the industry from incurred loss to expected loss. The new impairment model will lead to an increasing provision before the default occurs. In the short term, the IFRS 9 impairment model puts extra pressure on institutions and might prompt a shift from the standardized approach to the more challenging IRB one, and encourages banks to address their data governance shortcomings and break down the internal silos since building credit risk models especially need more internal data and need to use "reasonable and supportable information" to assess significant increases in credit risk. In the long term, the convergence between IFRS 9 and Basel III will improve risk management and bring greater integration with accounting practices. It will also provide stronger foundations for a more secure industry and improve confidence and transparency for all stakeholders in the market.

CHAPTER 3 METHODOLOGY

3.1 Introduction

Missing data has been a common phenomenon in SME data since often SMEs are not always willing or capable to report their internal data and they do not have to release data to the market if they are not listed companies. Yet, this problem could be getting resolved because the implementation of IFRS 9 for SMEs is beneficial to their accounting information to be more transparent and standardised.

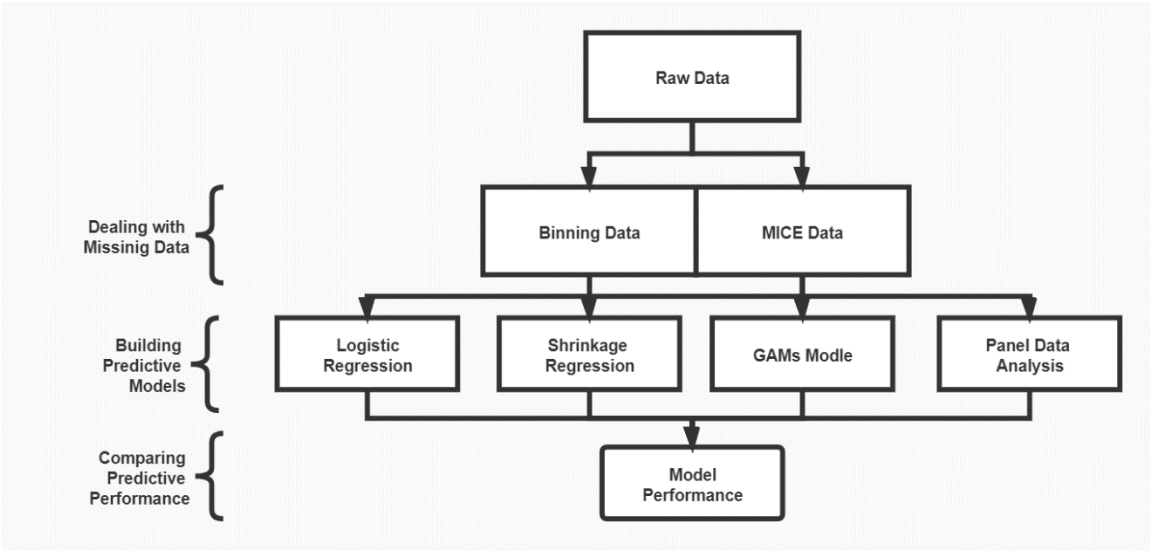


Figure 3-1 Development flow from raw data to model building.

Figure 3-1 provides an overview of the methodologies used in this research. Data cleaning is an essential step before data analysis and model building. However, only 1% of the observations are completely observed, and there is a need to deal with the missing data problem; otherwise, the analysis result would be inaccurate and unconvincing. The second and third sections introduce how to deal with missing data and how to select powerful predictors to look into the good/bad separation.

SMEs default prediction has always been a pivotal and popular topic in credit scoring as its impacts on bank lending and became even more crucial during the last credit crisis. Before extending a loan, banks need to judge the default probability of borrowers. In other words, one of the main objectives of credit scoring models is to assign credit applicants to either a ‘good credit’ group that is likely to repay financial obligation or a ‘bad credit’ group whose application may be denied because of its high possibility of defaulting on the financial obligation (T.-S. Lee, Chiu, Lu, & Chen, 2002). As a result, banks can make a better lending decision, so hopefully minimise loss and so hopefully a significant saving (Atiya, 2001). In addition, the risk associated with lending to small businesses has become more important since regulations started obliging banks to use separate procedures in assessing SMEs' credit worthiness (Andrikopoulos and Khorasgani, 2018). Basel Accord allows banks to use IRB approaches to determine minimum capital requirement, which has stimulated a great deal of interest to investigate the probability of default. This has been strengthened under IFRS 9 given the requirement to assess over a 12-month period and if there is change need to take account of the deterioration. Prediction is a typical classification problem where the objective is to determine which indicators are involved in success (or failure) of a corporation.

Despite the complexity of the prediction, a binary problem has usually been considered to tackle this classification problem (Alfaro, Gámez, & García, 2018). For this reason, this chapter illustrates different classification models applied to study changes in UK SMEs credit risk during the financial crisis after solving the missing data problem. Given that the SME data in this research is large enough, the percent of divided the training data and testing data has less influence on modelling compared with limited dataset. Followed by the most common ratio, 70% data is used to train the classifier, and the remaining 30% data is used to test the accuracy of the classifier models (M. Ma, 2016; Thottakkara et al., 2016).

This research compares four predictive modelling approaches: logistic regression, shrinkage regression, generalized additive model (GAMs), and panel data analysis. Logistic regression is a commonly used method in credit scoring literature and the predicted risk is either monotonically increasing or decreasing (Thottakkara, et al., 2016) and shrinkage regression is used to avoid overfitting problems because of penalty term. GAMs are additive regression models that can relax the monotonicity assumption of logistic models and offer advantage of estimating non-linear risk functions for continuous variables while still remaining interpretable compared with machine learning methods although they may provide a better predictive performance. The final consideration is to introduce a panel logistic regression, which attempts to add time effect for analysis by controlling the macroeconomic effect using either year dummy variables or macroeconomic variables. This research assesses each model's discrimination using the area under the AUC.

3.2 Multiple Imputation by Chain Equations (MICE)

Multiple imputation (MI) has been used in large datasets with thousands of observations and hundreds of variables (He, Zaslavsky, Landrum, Harrington, & Catalano, 2010). It operates under the assumption that the given variables used in the imputation procedure are missing data are miss at random (MAR) or missing complete at random (MCAR). Implementing MICE when data are not MAR could result in biased estimates. Regarding MICE, it is a special application of MI technique (Raghunathan, et al., 2001; S. van Buuren, 2007), and it is an alternative approach of joint modelling.

In the MICE procedures, a series of regression models are performed thus each variable with missing data is modelled conditional on the other variables in the data, which means each variable can be modelled according to its variable type. Table 3-1 below lists the model selection based on the types of variable. Difference MICE software packages differ in their exact implementation of this algorithm, but the general strategy remains the same.

Table 3-1 Summary of general MICE imputation models

Types of variable	The model used for imputation
Continuous variable	Linear regression or Predictive mean matching
Binary variable	Logistic regression
Ordinal variable	Ordinal logistic regression
Nominal variable	Multinomial logistic regression

To sum up, it is not yet known which imputation technique is most appropriate in which situation, and which is flexible and robust enough to work in a broad range of possible applications.

If there are a large number of variables with missing values, it is desirable to know the time spent on imputation because researchers have to carefully consider the computational time which is within an acceptable range if MICE is applied. Treiman (2014) found that approximately doubling the number of variables to be imputed increased the time to impute by a factor of four. A single iteration depends on the computer's processing speed. The speed of MICE procedure depends on both number and type of variables. For example,70 mostly categorical variables or 100 mostly continuous variables may pose difficulty; multinomial logistic regression takes noticeably longer than ordinary least squares or logistic regression (White, et al., 2011).

Imputation time can be estimated by multiplying the time for a single chain by both the number of burn-in iterations and imputations. When initially making model adjustments, it suggests using two imputations for testing the speed before increasing to a larger number of imputations for final models.

3.2.1 MICE steps

Let us assume there are four variables in the dataset: Y, X_1, X_2, and X_3 where Y is the dependent variable, X_1 is continuous variable with missing values, X_2

is a complete observed variable, and X_3 is a binary variable with missing values. X_1 has higher missing proportion than that of X_3. The MICE process can be broken down into the following steps (Azur, et al., 2011):

- Step 1: “place holder” imputation. A simple imputation, such as mean substitution for continuous variable or mode substitution for a categorical variable, is conducted for every missing variable in the dataset (X_1 and X_3);
- Step 2: The variable with the lowest missing proportion (X_3) is set back to missing because the imputation order is determined by the Missing rates. Overall, the variable with the lowest Missing percentage is still regarded as missing variable (X_3), and other variables are complete observed (Y, X_1, and X_2);
- Step 3: The variable with the lowest missing percentage is imputed by other variables. For example, logistic regression is performed as X_3 is a binary variable. The type of imputed variable determines the type of imputed model;
- Step 4: The missing values for X_3 are then replaced with predictive values from the logistics regression. Both observed and imputed values of X_3 is subsequently used as an independent variable in the following regression models of other variables;
- Step 5: Steps 2-4 are then repeated for each missing variable following the sequence. The cycling through each of the variables constitutes on iteration. At the end of one iteration, all of the missing values have been imputed with the prediction from regression models that depends on its variable type. Missing values being updated at each cycle. Repeating the imputation chain m times generate unique sets of imputed values;
- Step 6: Moving to the analysis phase. So far, there are M datasets, and each dataset performs the same statistical analyses and generates M different parameter estimates;
- Step 7: The final step is pooling phase. Pooling the M parameter values into a single point estimate (Donald B Rubin, 1987a), obtaining the MI parameter estimate and standard errors.

Each stage of the MI process is distinct and may be performed separately. Rubin's rules for combining estimate were, however, derived under the assumption that the imputation and analysis stages are conditioned on the same set of observed data. This implies that all variables included in the analysis stage should also include in the imputation stage, otherwise biased estimates would be produced.

3.2.2 Combining Rules

Donald B Rubin (1987a) proposed a series of rules to describe the combination of a single inference of multiple sets of parameter estimates and standard errors after the generation of a number of imputed datasets, and these rules also known as Rubin's Rules. MI parameter estimate is the arithmetic average of the m complete-data estimates, which mathematically is:

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i \tag{18}$$

where $\hat{\theta}_i$ is a parameter estimate from imputed dataset i and $\bar{\theta}$ is the pooled estimate. It should be mentioned that the foundation of MI is the Bayesian framework, but the pooled point estimate is valid for both a Bayesian and frequentist approach. On the one hand, $\bar{\theta}$ is the mean of the posterior distribution, on the other hand, $\bar{\theta}$ is a point estimate of the fixed population parameter (Donald B Rubin, 1987a) .

Pooling standard errors need to compute two components: within-imputation variance, and between-imputation variance. Within-imputation variance estimates the sampling variability that we would have expected had there been no missing data. The formula is given below:

$$V_W = \frac{1}{m}\sum_{i=1}^{m} SE_i^2 \tag{19}$$

where V_W denotes the within-imputation variance, and SE_i^2 is the squared standard error from imputed dataset i. This part is also simply the arithmetic mean of the sampling variance from each dataset.

Between-imputation variance quantifies variation in the parameter values caused by missing data, as follows:

$$V_B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2 \quad (20)$$

where V_B denotes the between–imputation variance, $\hat{\theta}_i$ is the parameter estimate from imputed dataset i and $\bar{\theta}$ is the average point estimate of parameter estimate from previthe ous equation. Again, between-imputation variance represents the additional sampling error that because of the missing data since the fluctuation of the $\hat{\theta}_i$ values came from the uses of different imputed datasets. Single imputation is sometimes considered as an alternative to MI, but it is unable to capture the between-imputation variance, hence standard errors are too small (White, et al., 2011).

Finally, the total sampling variance is a sum of the previous two components with an additional source of sampling variance, as follows:

$$V_T = V_W + V_B + \frac{V_B}{m} \quad (21)$$

where V_T denotes the total sampling variance. The additional source ($\frac{V_B}{m}$) represents the sampling error associated with the extra variance caused by the fact that coefficient estimates are based on finite m. It is used as a correction factor for using a specific number of imputations. When the number of imputation goes closer to infinity ($V_T = V_W + V_B$), the parameter estimate is more accuracy (Craig K Enders, 2010). The standard error is the square root of the total sampling variance, as follows:

$$SE = \sqrt{V_T} = \sqrt{V_W + V_B + \frac{V_B}{m}} \quad (22)$$

It is evident that standard errors should increase if missing values are imputed because these imputed values add an extra fluctuation to the parameter estimates, and this is why even the best single imputation technique, such as stochastic regression imputation, fails to estimate the correct standard errors.

3.2.3 Imputation Diagnostic Measure

The within-imputation variable, between-imputation, and the total variance define two useful diagnostic measure, the fraction of missing information and the relative increase in variance due to nonresponse. These measures are essential because they quantify the influence of missing data on the standard errors, determine the converge speed of the data augmentation algorithm, and help define the significance tests.

Missing information in an essential topic in statistical research on inferential methods for incomplete data (Todd E Bodner, 2008a). The fraction of missing information (FMI) estimates the missing data's influence on the sampling variance of a parameter estimate. It is estimated based on the percentage missing for a particular variable and how correlated this variable is with other variables in the imputation model. Longford (2006) has concluded that "information about a quantity is defined as the reciprocal of the mean squared error of its efficient estimator. The missing information about a parameter is defined as the difference of the information contained in the complete and incomplete datasets, and the fraction of missing information is the ratio of this difference and the complete-data information." In short, Paul D Allison (2001) stated that the fraction of missing information is "how much information is lost about each coefficient because of missing data". Both of them implies that missing information is specific for each parameter of interest and need not be equal for different parameters. Besides, the fraction of cases with missing values are not equivalent to the fraction of missing information, and the missing information is typically lower than the missing data rate, particularly when the variables in the imputation model are predictive of the missing data (Longford, 2006; Craig K Enders, 2010). The FMI formula is given below:

$$\mathrm{FMI} \approx \frac{\mathrm{V_B} + \frac{\mathrm{V_B}}{\mathrm{m}}}{\mathrm{V_T}} \tag{23}$$

The interpretation is similar to an R-squared. For example, an FMI of 0.15 implies that 15% of the total sampling variance is because of missing data. Provided that a variable with large proportion of missing values, the smaller FMI is, the more imputations are needed. The larger the number of imputations, the more precise the parameter estimates will be. The accuracy of the estimate of FMI increases as the number imputation increases because variance estimates stabilize with larger numbers imputations (Craig K Enders, 2010). A high FMI can indicate a problematic variable as high rates of missing information tend to converge slowly. If convergence of imputation model is slow, then it is necessary to examine the FMI estimates for each variable in imputation model. If FMI is high, then consider increasing the number of imputations. A good rule of thumb is to set the number imputations (at least) equal the highest FMI percentage.

Relative increase in variance (RIV) quantifies the proportional increase in total sampling variance that is due to missing information. The formula is given below.

$$\mathrm{RIV} = \frac{\mathrm{V_B} + \frac{\mathrm{V_B}}{\mathrm{m}}}{\mathrm{V_w}} = \frac{\mathrm{FMI}}{1 - \mathrm{FMI}} \tag{24}$$

Likewise, variables with substantial amounts of missing observations will tend to have higher RIV. An example to explain RIV is that RVI = 2.5 means that sampling fluctuation due to the missing data is two and half times larger than the sampling variance of sampling variability that we would have expected had there been no missing data. An extreme situation is that the missing data do not influence the sampling error of a particular parameter, the between-imputation variance is zero, as is the FMI and RIV.

FMI and RIV can indicate the convergence speed of the data argumentation algorithm, but they provide different aspects to view the within-imputation variance and between-imputation variance.

3.2.4 Number of Imputations

As mentioned above, MI may be unstable, and one of the greatest uncertainties in the practice of MICE is how many imputed datasets are needed to obtain reasonable and stable parameter estimates. Intuitively, a dataset with a large amount of missing information requires more imputations. The number of imputations needed is an important input as it could affect the accuracy of results. Relatively low values of number of imputations may still be appropriate when the fraction of missing information is low, and the analysis techniques are relatively simple.

Historically, the recommendation from standard texts was three to five imputed datasets, which based on the relative efficiency formula derived from RR (Donald B Rubin, 1987b). The relative efficiency (RE) of an imputation measures how well the true population parameters are estimated and is related to both the amount of missing information as well as the number of imputations performed. The formula is given below:

$$\mathrm{RE} = (1 + \frac{\mathrm{FMI}}{\mathrm{m}})^{-1} \tag{25}$$

where m is the number of imputation and FMI is fraction of missing information. RE is an estimate of the efficiency relative to performing an infinite number of imputations (m). As the number of imputations goes to infinity, FMI/m tends to be zero, and RE becomes 1. It may appear that you can get acceptable RE with a few imputations. When the amount of missing information is very low, then efficiency may be achieved by only performing a few. However, when there is a high amount of missing information, more imputations are typically necessary to achieve adequate efficiency for parameter estimates. An extreme example is if 10 per cent loss of efficiency is accepted, then RE is 90%, hence m = 9 if FMI = 90%. With even 90% of FMI, a small number of imputed datasets may be sufficient to keep majority of estimator efficiency compared to an infinite number of imputations. Thus, researchers can obtain relatively good efficiency even with a small number of imputations. However, this does not

mean that the standard errors will be well estimated since estimates of coefficients stabilize at much lower values of number of imputation than estimates of variances and standard errors. More imputations are often necessary for proper standard error estimation as the variability between imputed datasets incorporate the necessary amount of uncertainty around the imputed values.

Recently, the larger number of imputations are often recommended because of the rapidly developed computing power and practically used for researchers. Joseph L. Schafer and Graham (2002) found that 20 imputations can effectively perform better estimates by removing noise from other statistical summarizes (e.g., significant levels or probability values). John W Graham, Olchowski, & Gilreath (2007) approached the problem in terms of loss of power for hypothesis testing. Based on simulations and a willingness to tolerate up to a 1 per cent loss of power, they recommended 20 imputations for 10% to 30% missing information, and 40 imputations for 50% missing information. The recommended number of imputations is much larger than what is derived from the RE. A larger number of imputations may also allow hypothesis tests with less restrictive assumptions (i.e., that do not assume equal fractions of missing information for all coefficients). P. Allison (2012b) indicated other factors, such as standard error estimates, confidence intervals, and p-values need to be considered. One of the critical components of Rubin's standard error formula for MI is the variance of each parameter estimate across the multiple datasets. With so few observations (dataset), it should not be surprising that standard error estimates can be very unstable. Pan, Wei, Shimizu, & Jamoom (2014) pointed out that a small number of imputations may not be enough to obtain a statistically reliable variance estimate as well.

Since MI includes a random component, repeating the same analysis will give slightly different results each time unless setting up a seed of the random number generator to ensure the results are reproducible. This is obviously an undesirable property, but acceptable if the amount of variation is small enough

to be unimportant. The variation due to the random component is called the Monte Carlo error (MCE). White, et al. (2011) suggested the rule of thumb that the number of imputations should be at least equal to the percentage of incomplete cases and recommended the following guidelines for an acceptable amount of MCE[30]. Researchers should increase the number of imputations if conditions are not met:

- The MCE of a coefficient should be no more than 10% of its standard error;
- The MCE of a coefficient's t-statistic should be no more than 0.1;
- The MCE of a coefficient's p-value should be no more than 0.01 if the true P-value is 0.05, or 0.02 if the true p-value is 0.1.

Similar recommendations were proposed by (Todd E Bodner, 2008b), who also relied on simulation evidence, and by Royston et al. (2011), who analytically derived an approximation to the Monte Carlo error of the p-value. Despite their different approaches, and recommendations for the number of imputation, both sources agreed on the following simplified rule of thumb: the number of imputations should be slightly higher to the percentage of cases that are incomplete. For example, if 27% of the cases in the dataset have missing data, 30 imputed datasets should be generated.

Obviously, no single number of imputations will fit all situations. Therefore, specific guidelines for choosing number of imputation await empirical research. In general, it is a good practice to specify a sufficient number of imputation to ensure the converge of MICE within a reasonable computational time (Dong and Peng, 2013) as getting more datasets requires more time. With large datasets and many variables in the imputation model, this can become burdensome (P. Allison, 2012b). Multiple runs of a relatively large number of imputations are recommended to assess the stability of the parameter estimates.

[30] MICE packages in R do not provide MCE.

3.2.5 Non-normally Distributed Variables

Another uncertainty of MICE is specifying the imputation model correctly, especially for non-normally distributed variables. Non-normally distributed variables can be skewed, on limited-range or semi-continuous variables, which consist of a large proportion of responses with point masses that are fixed at some value and a continuous distribution among the remaining responses (Vink, Frank, Pannekoek, & van Buuren, 2014).

Researchers often treat non-normally distributed variables as general continuous variables and impute them under an assumption of normality using linear regression. The drawback of imputing such variables by assuming normality is that the distribution of imputed values does not resemble that of the observed values, and Von Hippel (2013) found that imputing skewed continuous variables under a normal model can lead to bias. Therefore, some researchers suggested to transform skewed variables to better approximate normality variables (J. L. Schafer and Olsen, 1998; Paul D Allison, 2001; Raghunathan, et al., 2001; Joseph L. Schafer and Graham, 2002). One option that is commonly used in practice is to apply a de-skewing transformation, such as the log or zero-skewness log transformation, prior to imputation (K. J. Lee and Carlin, 2010). However, one issue that arises with using a de-skewing transformation for positively skewed data is that when the imputed values are transformed back to the original scale, the imputed values can have very large outlying values (Von Hippel, 2013). Besides, the transformation does not yield normally distributed variable for both limited-range and semi-continuous data (White, et al., 2011).

Predictive mean matching (PMM) regression is a method of choice that imputes missing values of a continuous variable, especially for semi-continuous variable, such that imputed values are sampled only from the observed values of that variable by matching predicted values as closely as

possible, and it is very flexible as drawing imputations that relax some of the assumptions of parametric imputation, for example, it is free of distributional assumptions. The distribution of imputed variable often closely matches that of the observed variable, which means PMM tends to preserve the distributions of the original data, so the imputations remain close to the data. These properties generally appeal to applied researchers, but it is undesirable when the sample size is small since only a small range of imputed values is available (Heitjan and Little, 1991; Schenker and Taylor, 1996).

PMM calculates the predicted value using a regression model and picks the closest elements to the predicted value (by Euclidean distance). These chosen elements are called the donor pool (the observations potentially available for matching predictions), and the final value is chosen at random from this donor pool. The number in the donor pool is by default set to 5 in MICE (R packages). Thus the imputed value is an observed value whose prediction with the observed data closely matches the perturbed prediction (White, et al., 2011). Vink, et al. (2014) conclude that predictive mean matching performance is the only method that yields plausible imputations and preserves the original data distributions. If plausible values are necessary, this is a better choice than using bounds or rounding values produced from regression. Lazure (2017) compared five methods and showed that PMM under MICE has better performance in handling missing data.

3.2.6 Auxiliary Variable

A useful auxiliary variable is a potential cause or correlates of missingness or a correlate of the incomplete variables in the analysis model (Joseph L Schafer, 1997). Some researchers suggested to include as many variables as possible when doing multiple imputation (D. B. Rubin, 1996; Joseph L Schafer, 1997). Although there is no harm in using auxiliary variables with low (or zero) correlations, the most useful auxiliary variables are those that have correlations greater than positive or negative 0.40 with the incomplete analysis variables. Using auxiliary variables in a multiple imputation analysis is

particularly straightforward because the variables only play a role in the imputation phase. Including auxiliary variables in an analysis can improve the missing data handling procedure, either by reducing bias (i.e., better approximating the MAR assumption) or by increasing power (i.e., recapturing some of the missing information). Ideally, the auxiliary variables have no missing values, but this need not be the case.

3.3 Independent Variables Selection

3.3.1 Weight of Evidence and Information Value

As a popular tool in credit scoring, the use of information value (IV) and weight of evidence (WoE) has been in existence for more than 50 years (Weed, 2005). They have been applied as an effective tool to explore data and screen variables in credit scoring. When analyzing data, there are two distinct functions for IV and WoE. Weight of evidence describes the relationship between a predictive variable and the binary target variable, and IV measures the strength to predict the relationship between the dependent variable and an individual independent variable.

Since these two terms originate from credit scoring, they are always treated as a measure to separate good and bad customers. If a variable is described by continuous functions, then the following equation can calculate the information value:

$$\mathrm{IV} = \int \big(\mathrm{f(x|G)} - \mathrm{f(x|B)}\big) \log\left(\frac{\mathrm{f(x|G)}}{\mathrm{f(x|B)}}\right) \mathrm{dx} \qquad (26)$$

where $\mathrm{f(x|G)}$, and $\mathrm{f(x|B)}$ are the resulting conditional density function given a score x. However, it may difficult to estimate the density function sometimes, then an alternative approach to estimate IV of continuous variables is to make a discrete approximation to the density function. Rather than the integral over the continuous variables, this is done by splitting values into a number of bins and taking the summation over the bins (Thomas, 2009). The binning method is characterized by equal length of bins and tails are cut off to obtain a smooth

binning result. One thing that should be mentioned is that missing data of that variables is allocated to an individual bin so that it can be treated as a new attribute and enable to calculate WoE (G. Zeng, 2014). The IV thus can be calculated by the following formula:

$$\mathrm{IV} = \sum_{i=1} \left(\frac{g_i}{\sum_{i=1} g_i} - \frac{b_i}{\sum_{i=1} b_i} \right) \times \log \left(\frac{\frac{g_i}{\sum_{i=1} g_i}}{\frac{b_i}{\sum_{i=1} b_i}} \right) \tag{27}$$

where b_i and g_i are numbers of bad customers and good customers falling into a characteristic group. This formula is able to compute IV of categorical variables as well. One of their features is that they can assist in variables selection for both categorical and continuous. Siddiqi (2012) suggested that variables with extremely high information value (say greater than 0.5) are suspicious, and he provided a rule of thumb for selecting predictive variables: if information value is between 0.1 and 0.5 then it can be accepted as a predictor, and those variables with information values less than 0.1 are weak even useless predictors. Hence, discarding those variables with IV less than 0.1 is able to limit the scale of the analysis. Yet the number of independent variables could be still larger than that of a robust model.

On the other hand, WoE shows the different performance of groups and conveys information on the relative risk associated with each category of the particular variable, with a large negative value indicating a higher risk of default. It is mathematically defined as followed:

$$\mathrm{WoE}_i = \log\left(\frac{g_i / \sum_{i=1} g_i}{b_i / \sum_{i=1} b_i}\right) \tag{28}$$

Traditionally, when using an indicator for each class, it needs to estimate multiple regression parameters or dummy variables. This may introduce considerable additional variance into the model and make it unnecessarily complicated, but one only needs to estimate one regression parameter for a categorical variable using WoE. More importantly, it takes account for handling missing data by taking missing data as an individual attribute. One point should be mentioned is that traditionally, it is inappropriate to use IV to select variables to build a non-binary classification model, especially when there is an

insufficiently large number of observations, the denominator of WoE formula for both ‘good’ or ‘bad’ is probably going to be zero.

As mentioned, Weight of Evidence (WoE) is a popular tool in the fields of credit scoring to deal with the missing data. First, variables transformed by WoE and those variables with information value (IV) greater than 0.1 are said to have medium predictive power. Next, multiple imputation is employed to fix the missing data problem. After that, pooled estimate results and the Wald’s test are repeatedly used to determine the predictive variables.

3.3.2 Analysis with Multiple Datasets (after Imputation)

The analysis process follows once the imputation process has finished. There are a number of imputed dataset (m datasets), and the one of the most straightforward approaches to conduct analysis is to use Rubin’s rule to combine the estimates from each imputed dataset. Yet this is limited by incorrectly specified F distribution and overfitting (Zhao and Long, 2017). As mention in the last chapter, compared with single imputation, one of the strengths of multiple imputations is that standard error would be more accurate because of considering the between-imputation variance and the extra variability from a number of imputations when pooling multiple imputed datasets by Rubin’s rules.

Another approach is to combine all imputed dataset as a “complete” dataset and conduct analyses. However, it would theoretically weaken the advantage if running a regression analysis with a stacked dataset without any adjustment. Angela M Wood, White, & Royston (2008) proposed a stacked method and weighting scheme considering the fraction of missing data information in each covariate in case the standard errors are underestimated. A premise is that missing data do not exist in both dependent and independent variables given missing data are MAR or MCAR. Once multiple imputed datasets are obtained, one may stack all imputed datasets as a single large complete dataset. In general, the estimates based on the stacked MI data are unbiased if the

estimates based on a single dataset are unbiased, while the standard errors based on the stacked MI data will be under-estimated if they can be estimated (J. Cohen, Cohen, West, & Aiken, 2013). A simple way to correct the underestimated standard errors is to apply a weight to each individual (Angela M Wood, et al., 2008). Denote this weight by w_i for subject i, and there are three possible sets of weights in the following:

- First, one could assign $w_i = 1/M$ thus the overall weight for a subject is 1. This weighting scheme puts the same weight for each subject and ignores the degree of missing information.
- Second, $w_i = (1-f)/M$, where f, the average fraction of missing data across all variables, is calculated as (total number of missing values across all variables) divided by (total number of variables times total number of observations).
- Third, $w_i = (1-f_i)/M$ where f_i, the fraction of missing data for variable X_i, is calculated as (number of missing values of variable X_i) divided by total number of observations.

More recently, Wan, Datta, Conklin, & Kong (2015) applied similar weighted strategy on the stacked imputed datasets as well. They propose to assign the weight $w_i = f_i /M$, where f_i is the fraction of observed values for subject i, i.e. the ratio of number of observed variables for the subject i to the total number of predictor variables.

A variable with more missing information should be assigned less weight. Putting different weights on observation is obviously more appropriate way as it assigns weights according to the quality of the observed information, but it is unfeasible due to a lack of software support. Same weight according to fraction of missing information on all observations would be more workable in practice.

3.3.3 Empirical

A previous PhD thesis (M. Ma, 2016) empirically determined a set of variables presented in Appendix A[31]. First, the author employed WoE to transform variables. Then the author repeatedly used stepwise logistic regression for variable selection until all selected variables presented positive coefficient with confidence level 95% for each year as a negative coefficient would imply collinearity. Finally, in order to obtain the same set of predictive variables, and reduces annual variation through whole observed periods, the author decided to select those variables that are significant at least three years. It can guarantee that the number of predictors is in a manageable size and keeping variables being insignificant in one year is because those variables capture annual difference.

From Table 3-2, both start-ups and non-start-ups SMEs has three predictive variables with over 70% missing observations, especially, time since last derogatory data item (months) with 96.69%. Both start-ups and non-start-ups SMEs are unwilling to report their Proportion of current directors to previous directors in the last year, time since last annual return and time since last derogatory data item (months). In addition, start-ups are evasive about their SIC code and total assets, and debt gearing (%) for non-start-ups. Eventually, the percentage of complete observation is less than 1%. This means that complete-case analysis is not appropriate since it may generate a perfect prediction problem (dependent variable of all complete cases are classified as ‘good’ credit).

[31] The dataset used in this thesis is the subset from (M. Ma, 2016) PhD thesis.

Table 3-2 Missing rates (%) in the dataset

Variables	Start-ups				Non-start-ups			
	2007	2008	2009	2010	2007	2008	2009	2010
Legal form	0	0	0	0	0.21	0.07	0.01	0.03
Company is subsidiary	0	0	0	0	-	-	-	-
Parent company – derog details	-	-	-	-	0	0	0	0
1992 SIC code	58.39	60.21	58.53	60.93	3.54	3.59	4.26	3.58
Region	1.51	1.4	1.39	1.41	2.78	2.52	2.05	2.33
Proportion of current directors to previous directors in the last year	81.54	85.52	87.87	88.79	93.76	94.05	95.48	92.52
No. Of 'current' directors	-	-	-	-	2.12	1.98	1.81	1.9
Oldest age of current directors/proprietors supplied (years)	15.46	9.36	3.38	1.41	-	-	-	-
Number of directors holding shares	0.72	0.64	0.84	0.6	-	-	-	-
Pp worst (company DBT - industry DBT) in the last 12 months	-	-	-	-	66.81	69.27	65.26	66.86
Total value of judgements in the last 12 months	0	0	0	0	0	0	0	0
Number of previous searches (last 12m)	0	0	0	0	0	0	0	0
Time since last derogatory data item (months)	96.69	95.04	90.21	92.09	96.35	93.74	90	86.83
Lateness of accounts	0.47	0.51	0.52	0.6	1.69	1.55	1.41	1.46
Time since last annual return	57	56.93	52.8	56.38	2.73	2.39	2.92	2.43
Total assets	79.1	80.17	74.34	74.9	-	-	-	-
Total fixed assets as a percentage of total assets	-	-	-	-	3.75	4.33	4.69	4.75
Debt gearing (%)	-	-	-	-	92.03	91.93	94.21	91.74
Percentage change in shareholders' funds	-	-	-	-	8.09	8.71	8.76	8.71

Notes: The Variables with missing rates over 80% have been underlined.

3.4 Logistic Regression

Logistic regression (Cox, 1958) is the classical statistical techniques for credit risk modelling because of its ability to model binary classification problem (Andreeva, Calabrese, & Osmetti, 2016). In theory, it is supposed that logistic regression is a more proper statistical method for classification than linear regression due to its binary classes of the dependent variable (good and bad risk). Logistic regression provides a best linear fit using maximum likelihood method so it maximised the probability of allocating the observation into suitable category given the regression coefficients. Given its strong theoretical support, it gives rise directly to an additive log odds score which is a weighted linear sum of attribute values (Thomas, 2009). Wiginton (1980) published one of the first papers using logistic regression applied to credit scoring compared with the discriminant analysis. The authors summarized that in comparison with other methods, logistic approach provided a better classification result. As a standard benchmark, newly developed classifier algorithms compare classification performance against it (T.-S. Lee, et al., 2002; Ong, Huang, & Tzeng, 2005; Bellotti and Crook, 2009; Nehrebecka, 2018), and there are instance of developed logistic regression: piecewise logistic regression (R. Anderson, 2015), and fuzzy logistic regression (Sohn, Kim, & Yoon, 2016).

Let assume that $x = (x_1, x_2, \ldots, x_p)$ are a vector of p independent variables and y is a binary variable. Assume there is a sample of N independent observations (yi, xi1, xi2, ..., xip), for i = 1, 2, ..., N, where yi is the value of y (1 for bad customers and 0 for good customers) and xi1, xi2, ..., xip are the value of x1, x2, ..., xp for the ith observation. The logistic regression model is given by the equation:

$$P(y_i = 1|x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}} \qquad (29)$$

where β_0 is the intercept and $\beta_1, \ldots, \beta_p$ is coefficients of variable $x_1, \ldots, x_p$.

Specifically, logistic regression will fit a linear regression equation of predictors to predict the logit transformed binary (Good or Bad) dependent on variable Y, and the formula is given below.

$$\log\left(\frac{P(y_i = 1|x)}{P(y_i = 0|x)}\right) = \log\left(\frac{P(y_i = 1|x)}{1 - P(y_i = 1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (30)$$

Regarding the model fit, two hypotheses are of interest:

- H_0: all the coefficients in regression equal zero, and
- H_A: the predictors have a significant impact on prediction.

The likelihood ratio test is based on -2log likelihood ratio, and it is a test of the significance of the gap between the likelihood ration (-2log likelihood) for models with predictors minus the likelihood ratio for baseline model (constant only). Chi-square is used to determine significance of this ratio.

Independent variables transformed by weight of evidence are particularly well suited for logistic regression because such a transformation allows maintenance of linear dependence in regard to the logistic function. The link between logistic regression and weight of evidence is provided in the following equation. Besides, they also have ties to well-known naive Bayes classifier, given by (Jerome Friedman, Hastie, & Tibshirani, 2001):

$$\log\left(\frac{P(y_i = 1|x_1, \ldots, x_p)}{P(y_i = 0|x_1, \ldots, x_p)}\right) = \log\left(\frac{P(Y = 1)}{p(Y = 0)}\right) + \sum_{i=1}^{n}\left(\frac{f(x_n|Y = 1)}{f(x_n|Y = 0)}\right) \quad (31)$$

The left-hand side of the equation above, the conditional log odds are precisely the logit transformation. The first term of the right-hand side is log odds and is a constant, and the second term is Weight of Evidence in the form of continuous variables. This is also why the greater value of Weight of Evidence, the higher chance of observing Y = 1. This equation comes from assuming that all predictors are conditionally independent given Y, which is a highly optimistic assumption (Larsem, 2015). Therefore, a “semi-Naive” version of this model is introduced.

$$\log\left(\frac{P(y_i = 1|x_1, \ldots, x_p)}{P(y_i = 0|x_1, \ldots, x_p)}\right) = \log\left(\frac{P(Y = 1)}{p(Y = 0)}\right) + \sum_{i=1}^{p}\beta_j\left(\frac{f(x_p|Y = 1)}{f(x_p|Y = 0)}\right) \quad (32)$$
$$= \beta_0 + \beta_1 WoE_1 + \cdots + \beta_p WoE_p$$

The idea is to transform the data into weight of evidence vectors and then use logistic regression to fit the model. Hence, partly relaxing the assumption that all predictors in the model are independent.

3.5 Shrinkage Regression

There is a trade-off between bias and variance regarding establishing a prediction model. Bias simply means how far away is estimated values from actual values, and variance is a measure of spread or variations in predictions. High bias can cause an algorithm to miss the relevant relations between independent variables and the dependent variable, while high variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs.

Traditionally, stepwise selection, a common used approach to select useful variables, may produce an interpretable model and has possibly lower prediction error than the full model by discarding a subset of the predictors. Yet, since it is a binary process, which means variables are either kept or discarded thus, it often presents high variance. Shrinkage methods, however, are more continuous and would not suffer such high variability as stepwise selection. It fits a model including all variables and using constrains to regularize the coefficient estimate. Classical estimates are unconstrained, and they can have high value for the coefficients, which would result in very high variance in the model leading to overfitting. Shrinkage method shrink coefficients towards zero by adding a constraint. As a consequence, the variance of model reduces by estimating some of the coefficients to be zero, and so variable selection is performed (Jerome Friedman, et al., 2001; James, Witten, Hastie, & Tibshirani, 2013).

The two best-known shrinkage approaches are ridge regression and lasso regression. Regarding ridge logistic regression (Hoerl and Kennard, 1970), a penalized parameter applied all the coefficient estimates except the constant. Ridge regression is known to shrink the coefficients of correlated predictors towards each other, allowing them to borrow strength from each other (J. Friedman, Hastie, & Tibshirani, 2010). Tibshirani (1996) proposed a shrinkage method ‘least

absolute shrinkage and selection operator', also known as lasso, and the author has defined lassos from linear regression to other regression methods, including a lasso for logistic regression achieved by replacing the residual sum of squares by the corresponding negative log-likelihood function (Meier, Van De Geer, & Bühlmann, 2008). Lasso performs both variable selection and regularization in order to improve the prediction power and interpretability of the regression model. Estimates of coefficient are sparse, which indicates that coefficient of some variables may be exactly zero. This automatically removes irrelevant variables.

Figure 3-2 from (Jerome Friedman, et al., 2001) illustrates the lasso (left) and ridge regression (right) when there are only two parameters so that the plot is on a surface (if more than two variables, the plots would be on a space or even hyperplane), and this is beneficial to understand why lasso can perform variables selection. The residual sum of squares has elliptical contours, centered at the full least squares estimate.

The constraint region for lasso is the square $|\beta_1| + |\beta_2| \leq t$, while that for ridge is the circle $\beta_1^2 + \beta_2^2 \leq t$. If t is sufficiently large, then the constraint regions will contain $\hat{\beta}$, and so the ridge regression and lasso estimates will be the same as the least squares estimates.

On the other hand, if it is sufficiently small that least squares estimates lie out the constrain areas, both methods find the first point as the residual sum of squares expands where the elliptical contours hit the constraint region. Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will not likely to be zero. Unlike the ridge constrain, the lasso constrain has sharp points; if the ellipse intersects the constrain region at an axis (corner) then one parameter β_j will equal to zero. At higher dimension (number of variable increases), the diamond becomes a rhomboid, and has many corners, flat edges and faces; there are many more opportunities for the estimated parameters to be zero.

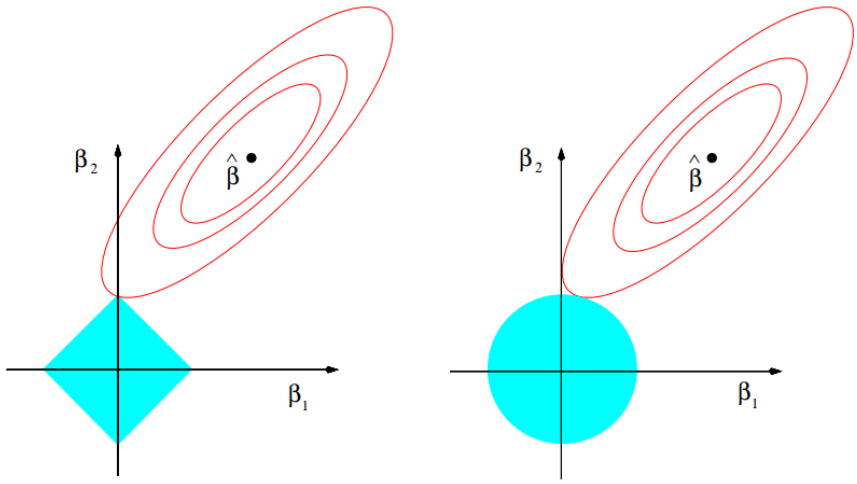


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Figure 3-2 Figure that help explains why lasso can select predictors, taken from Jerome Friedman, et al. (2001)

Given then the logistic regression:

$$\log\left(\frac{P(y_i = 1|x)}{P(y_i = 0|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0 + x^T\boldsymbol{\beta} \tag{33}$$

where β_0 denotes the intercept, $\beta = (\beta_1, \cdots, \beta_p)$ denotes the linear coefficients. So, the desire is to determine the coefficients (β_i) that makes SMEs labelled as 'good' has a value as close as possible to one and the 'bad' has a value as close as to zero, thus a distinct gap between 'good' and 'bad' SMEs' credit. A maximum likelihood approach is a typical way to achieve this goal, and the log-likelihood transformation makes it more accessible to compute mathematically. The log-likelihood function can be written:

$$\begin{aligned} l(\beta_0, \beta) &= \sum_{i=1} y_i \log p(y_i = 1) + (1 - y_i)\log\big(1 - p(y_i = 1)\big) \\ &= \sum_{i=1} y_i\big(\beta_0 + x_i^T\beta\big) - \log\left(1 + e^{\beta_0 + x_i^T\beta}\right) \end{aligned} \tag{34}$$

Moreover, minimizing the negative log-likelihood is the same as maximizing the log-likelihood. The log-likelihood function can be further extended by adding penalty of the L_1 penalty $\sum|\beta_i|$ (lasso) or the L_2 penalty $\sum \beta_i^2$ (ridge). Besides, Zou and Hastie (2005) introduced an approach, called elastic net, includes a non-negative tuning parameter α, being the penalty a mixture of the previous two approaches (lasso and ridge). Elastic net is particularly useful when the number of predictors is much larger than the number of observations (J. Friedman, et al., 2010). Finally, the objective function for the penalized logistic regression of Lagrange form applied the negative binomial log-likelihood, and is:

$$\min_{(\beta_0,\beta)} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \left(\beta_0 + x_i^T\beta\right) - \log\left(1 + e^{(\beta_0 + x_i^T\beta)}\right)\right] + \lambda\left[\frac{(1-\alpha)\|\beta\|_2^2}{2} + \alpha\|\beta\|_1\right] \quad (35)$$

where $\lambda \geq 0$. When α = 1 refers to pure lasso, α = 0 refers to pure ridge, otherwise elastic net. As λ increases, the flexibility of fit decreases, leading to decreased variance but increased bias.

Determining λ is a significant step for both ridge and lasso as they generate a set of coefficient estimates whose values depend on the various values of λ. One useful way for tuning parameter λ is known as cross-validation (Jerome Friedman, et al., 2001), which is an approach of assessing how well a model can be generalized to an independent dataset. In general, the best lambda ( ) for the data, can be defined as the lambda that minimize the cross-validation prediction error rate. K-fold cross-validation splits the data into k subsets of around equal size and one of the subsets becomes the validation set. The remaining k-1 subsets are used as training data. This procedure is repeated k times, each time with a different validation set, and the optimum value of λ is estimated and hence the cross-validated log-likelihood is maximized (Pereira, Basto, & Silva, 2016). Cross-validation can be used to select α as well, although it is often viewed as a higher-level parameter and chosen on more subjective grounds.

To perform the lasso and ridge regression, there is one package on R called glmnet (Jerome Friedman, Hastie, & Tibshirani, 2009). The package is able to fit generalized linear model with L_1 or L_2 regularization even elastic net. In addition, the package provides two lambda. Regarding linear regression, one lambda (lambda.min) will make the error minimum, while another lambda (lambda.1se) will make the error within one standard error. As suggested by (Jerome Friedman, et al., 2001; James, et al., 2013), lambda.1se is preferred when selecting the best model.

Ridge regression and lasso perform by trading off a small increase in bias for a large decrease in variance of the predictions. Hence they may improve the overall prediction accuracy (Pereira, et al., 2016) since lasso regression may involve only a subset of the predictors, which in turn improves model interpretability. Considering prediction accuracy, one can expect that lasso to perform better generally when only a small number of predictors have substantial coefficients, while when all coefficients are roughly of equal size, one expects a better performance of ridge regression (James, et al., 2013).

3.6 Generalized Additive Models (GAMs)

It is not surprising that both logistic regression and shrinkage regression are prevalent and widely used because of their established advantages. Yet, the simplicity of the model is achieved by assuming that there is a linear relationship between dependent variables and independent variables. This is, however, an oversimplification of the real relationship within the data. If the relationship is not correct, then the estimates of the coefficients and the inferences based on them could be misleading (P. D. Allison, 1999; Jerome Friedman, et al., 2001; Horowitz and Savin, 2001).

In addition, Interpretable models, though sometimes less accurate than blackbox models, are preferred in many real-world applications and they are often used because their transparency helps determine if a model is biased or unsafe(J. Zeng, Ustun, & Rudin, 2017; Tan, Caruana, Hooker, & Lou, 2018). Generalized additive

models (GAMs) are among the most powerful interpretable models when individual features play major effects (Lou, Caruana, & Gehrke, 2012).

GAMs proposed by (T. Hastie and Tibshirani, 1986) is a flexible statistical approach which can identify and capture non-linear regression effects, which are in turn an extension of the classical linear model, and maybe more close to the real relationship. The GAMs approach has distinct advantages. GAMs do not involve strong assumptions about the relationship between two or more variables that is implicit in standard parametric regression. Such assumptions may force the fitted relationship away from its natural path at critical points. Moreover, it allows automatic fitting of a non-linear function to each independent variable so that it is possible to provide substantial new insights into the effects of the independent variables. This means that it is unnecessary to explore the non-linear relationship on each variable individually by manual transformations (James, et al., 2013). Since there is a more complex relationship, it is likely to obtain a more accurate prediction (T. Hastie and Tibshirani, 1986; James, et al., 2013). In this research, GAMs use a sum of smooth functions and have successfully proven their ability to capture non-linear relationships between independent variables and a dependent variable in many areas (Dominici, McDermott, Zeger, & Samet, 2002; Austin, 2007; Berg, 2007; Aalto, Pirinen, Heikkinen, & Venäläinen, 2012; M. Ma, 2016). Therefore, GAMs becomes an attractive alternative to logistic regression to explore SMEs performance from the dataset (T. Hastie and Tibshirani, 1986, 1987).

Additive logistic regression can be modelled in a non-parametric way, and the data decides on the functional form. Given the logistic regression:

$$\log\left(\frac{\mathrm{P(X)}}{1-\mathrm{P(X)}}\right) = \beta_0 + \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}}' \mathrm{X_j} \qquad (36)$$

where P(X) = P(Y = 1|X), β_0 denotes the intercept, and β_{j}' = (β_1, ···, β_{p}) denotes the linear coefficients. The additive logistic regression model replaces each linear term $\sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}}' \mathrm{X_j}$ by a more general functional form:

$$\log\left(\frac{P(y_i = 1|x)}{P(y_i = 0|x)}\right) = \beta_0 + \sum_{j=1}^{p} s_j(X_j) \quad (37)$$

Where s(x) denote a smooth function. Notice that smooth functions cannot be applied to non-continuous variables.

Smoothers are central to GAMs. A smoother is a mathematic technique for approximating an observed dependent variable by a smooth function of one (or several) independent variable(s). Smoothers, such as smoothing splines and regression splines are non-parametric because they make no parametric assumption about the shape of the function being estimated. In general, the amount of smoothing selected will have more impact on the final function than the type of smoother chosen (Ramsay, Burnett, & Krewski, 2003).

Polynomial based smooth function is widely used in non-linear modelling, yet as the number of basic functions increases, the polynomial bases become increasingly collinear. The larger the number of basic functions is, the 'wigglier' the non-linear estimate (smooth term) is. This yields highly correlated parameter estimators, and leads, as a consequence, to high estimator variance as well as numerical problems (S. N. Wood and Augustin, 2002). For these reasons, adding several polynomial terms does not represent a valid solution to capture non-linear relationships.

Splines are used in order to overcome these issues, because it is the sums of weighted basis functions which have convenient mathematical properties and good numerical stability, and the properties of basic functions rely on the type of splines. The complexity and flexibility of the fit depend on the number of basic functions. Thus, spline bases could be employed to determine the relationship between the continuous predictors and the outcome of interest.

Common choices for representing smooth functions include natural splines and smoothing splines (Wahba, 1990; Green and Silverman, 1993; T. J. Hastie, 2017). However, the problem is that a spline basis can be constructed only if using knots

at fixed locations throughout the range of the data. In particular, the choice of knot locations introduces some subjectivity into the model fitting process which may result in a substantial effect on the resulting smooth. Smoothing splines avoid this problem by placing knots at every data point and are indeed sometimes referred to as full rank smoothers because the size of the spline basis is equal to the number of observations. This, however, comes at a cost, the iterative smoothing parameter estimation can be computationally heavy, especially with large datasets (Leathwick, Elith, & Hastie, 2006), and such smoothers have as many unknown parameters as there are data and hence the difficulty is computational cost.

S. N. Wood (2003) proposes using thin plate regression splines (TPRS) smoothing function because it is a low rank smoother such that no need to select knot locations, and reasonably increasing computational efficient. Besides, TPRS has similar performance of a full rank TPRS. TPRS are constructed by starting with the basis and penalty for a full thin plate spline and then truncating this basis in an optimal manner, to obtain a low rank smoother. A TPRS smoothing function with a various number of basic dimensions at each independent variable is used. Details are given in (S. N. Wood, 2003). One key advantage of the approach is that it avoids the knot placement problems of conventional regression spline modelling, but it also has the advantage that smooths of lower rank are nested within smooths of higher rank so that it is legitimate to use conventional hypothesis testing methods to compare models based on pure regression splines. It is necessary to create all the basic functions before doing the truncation or decomposition of them, even though the truncation process allows for far fewer basis functions to be used in fitting (S. N. Wood, 2003).

This thesis, therefore, uses regression splines to estimate the non-linear trend. In addition, Berg (2007) indicated that since each of the individual additive terms are estimated using univariate smoothers, GAM avoids the problem of rapidly increasing variance for increasing dimensionality. This problem is referred to as the ‘curse of dimensionality’ and is present in many non-parametric methods. Specifically, the regression spline of an independent variable is made up of a linear combination of known basis functions, $b_{jk}(x_j)$, usually chosen to have good

approximation theoretical properties, and unknown regression coefficient parameters, δ_{jk},

$$s_j(x_i) = \sum_{k=1}^{q_j} \delta_{jk} b_{jk}(x_j), \tag{38}$$

Where j indicates the smooth term for the jth independent variable, qj is the number of basis function, and hence regression parameters used to represent the jth smooth term. With each s_j is associated a with smoothing penalty, which is quadratic in the basis coefficients and measures the complexity of s_j. Writing all the basis coefficients in one p-vector $\boldsymbol{\beta}$, then the jth smoothing penalty can be written as $\boldsymbol{\beta}^T \mathbf{S}^j \boldsymbol{\beta}$, where $\mathbf{S}^j$ is a matrix of known coefficients, but generally has only a small non-zero block. The estimated model coefficients are then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ -l(\beta) + \sum_{j}^{M} \lambda_j \boldsymbol{\beta}^T \mathbf{S}^j \boldsymbol{\beta} \right\} \tag{39}$$

given M smoothing parameters, λ_j, controlling the extent of penalization (S. N. Wood, Pya, & Säfken, 2017). Trevor, Robert, & JH (2009) showed that a small number of degrees of freedom (df = 4) well fits most dataset.

In order to optimize and estimate the GAM model, mixed model approach via restricted maximum likelihood (REML) is applied. Technical details can be found in S. N. Wood (2004). The basic idea is to recast a GAM as a parametric, penalized GLM. They also showed that REML penalizes overfitting more and therefore have more pronounced optima, leading to fewer optimisation issues and less variable estimates of the smoothing parameter (S. N. Wood, 2011). The package in R that used to fit GAMs is mgcv written by S. Wood and Wood (2015). Marra and Radice (2010) concluded that the mgcv package are preferred to the use of the gam library. S. N. Wood (2003) indicated that broadly speaking the default penalized TPRS in the package tends to give the best MSE performance, but they are slower to set up than the other bases because of placing the knots.

Regarding smoothness, its levels are about choosing values for the λ_j (S. N. Wood, 2008). Each smoother has a parameter that determines how smooth the resulting

function will be. Instead of providing the output of λ_j, the mgcv package uses a term called the effective degrees of freedom (EDF) as a consistent way to quantify model, which can be interpreted as smoothed level of a variable (the higher the EDF, the more non-linear and complex the splines). The number of free parameters in GAMs is difficult to define, the EDF are instead related to λ_j, where the effect of the penalty is to reduce the EDF. EDF decreases as the penalty λ_j increases until the best fit penalty is found. Increasing the EDF would make the smooth term wriggle more, and then an explanation of the trend may become difficult.

3.7 Panel Data Analysis

Panel data (longitudinal data) analysis is a means of studying an entity who are surveyed periodically over a given time span. Using repeated observations of sufficient entities, panel analysis allows the researcher to explore the change. The combination of time series with cross-sections can boost the quality and quantity of data in ways that would be impossible using only one of these two dimensions (Gujarati, 2014). Besides, these longitudinal data have more variability and allow researchers to investigate more issues than do cross-sectional or time-series data alone (Kennedy, 2003).

The natural differences of observed components such as legal form, industry sectors, and regions across UK, as well as unobserved components suggest that SMEs have their own firm-specific characteristics. Analysis of panel data is able to control for individual heterogeneity which is unobservable while both cross-section and time-series study cannot (Baltagi, 2008). For example, this research is to explore SMEs performance during the credit crisis, and its performance is modelled by some measurable variables recorded in the dataset. However, there are still a number of other variables that may change only across firms or across times. Examples for the time-invariant variables are educational level of firm's leaders, risk appetite of decision makers. Examples for the firm-invariant variables are national policies and international agreements. These unobservable

(unrecorded) variables may have impact on SMEs' performance more or less. Omission of these variables results in biased estimates.

Additionally, it is likely that during the last credit crisis, there were macroeconomic shocks and changes in the institutional context. Cross-sectional analysis on SMEs is based on an annual view, thereby failing to capture the time series effect. Given that the performance of observation at the peak of credit crisis (2009) may differ from observation in normal period (2007) or recovery period (2010), time effect is of interest to study. Moreover, previous methods are based on firm-specific variables only, which means there is no correlation between SMEs' performance and business cycle. This does not hold for reality.

For these purposes, panel data analysis is available to provide a different aspect to find out the truth behind the dataset, and is possible to include time effects as well as to control for the heterogeneity of SMEs and reduce collinearity among the variables (Hsiao, 2014). Likewise, panel data analysis enables researchers to eliminate the potential omit variable biases in the resulting estimates because of correlation between unobservable individual effects and the independent variables in the prediction model (Michaelas, Chittenden, & Poutziouris, 1999). There are a number of studies exploring SMEs' performance using panel data. Nunes and Serrasqueiro (2012) used probit regression to estimate the survival determinants of young and old Portuguese SMEs considering the scale effect, financial characteristics, technological intensity, and macroeconomic situation using data from 1999 to 2006. (W. L. Lin, Yip, Sambasivan, & Ho, 2018) showed that Malaysian SMEs face financing problems as well. They used Generalised Method of Moment (GMM) with panel data to analyse the determinants of capital structure and found that firm size and asset structure have a significantly positive effect on the leverage ratio in SMEs, while age and taxation have a negative effect.

Panel analysis can provide a rich and powerful study of entities, if one is willing to consider both the space and time dimension of the data, which is knowns as two-way error component error regression model but it is difficult and complex to estimate and analysis. The panel data model may be represented using logistic

regression form due the nature the binary dependent variable and the popularity in the industry field:

$$\log\left(\frac{P(y_{it}=1|x_{it})}{P(y_{it}=0|x_{it})}\right)=\alpha+x_{it}\beta+v_{it}=\ \alpha+x_{it}\beta+(\mu_i+\lambda_t+\varepsilon_{it}) \qquad (40)$$

where y_{it} is the dependent variable, with i denoting entity (cross-section dimension) and t denoting years (time series dimension) ranging from 2007 to 2010. To model individual heterogeneity, one usually assumes that the error term (v_{it}) has three separate components, μ_i is specific to the individual and does not change over time, λ_t is time effect and ε_{it} is a random disturbance term of mean 0, and usually assumed well-behaved and independent of both the regressors x_{it} and the individual error component (Croissant and Millo, 2008).

However, as this research plans to explain the time only effect by introducing macroeconomic variables, only time-invariant individual-specific effects are considered. Therefore, the model becomes one-way error component regression model.

$$\log\left(\frac{P(y_{it}=1|x_{it})}{P(y_{it}=0|x_{it})}\right)=\alpha+x_{it}\beta+v_{it}=\alpha+x_{it}\beta++\mu_i+\varepsilon_{it} \qquad (41)$$

3.7.1 Fixed versus Random Effects

The treatment of individual-specific effects introduces two research methods: fixed effect (FE) and random effects (RE). It assumes that there is unobserved heterogeneity across individual captured by μ_i. The main difference is whether the individual-specific effects μ_i are correlated with the regressors. FE assumes that individual-specific effects is correlated with the regressors while RE assumes that individual-specific effects are distributed independently of the regressors. Wooldridge (2010) further indicated the discussion between FE and RE focus on whether μ_i is properly regarded as a random variable (RE) or as a parameter to be estimated (FE).

Selection between FE and RE for nonlinear models is considered more important than that for linear models since the choice will impact the analysis and lead to significant differences to estimated results. Baltagi (2008) indicated that the selection between FE and RE has generated a hot debate in many fields. This section discusses the logit panel data model and estimators selection.

FE investigates the relationship between independent variables and the dependent variables within an entity SME. Each SME has its own individual characteristic that may or may not has impact on the independent variables or dependent variables (good/bad). For example, risk appetite of an SME may influence its performance. This is because there is a correlation between the error term and independent variables. In order to control for this, FE is able to get rid of the effect of those time-invariant variables, so it is available to explore the pure effect between independent variables and dependent variables. Time-invariant variables are unique to the individual entity and should not correlated with other individual entity.

FE examines differences between SMEs in intercepts by assuming the same slopes and constant variance across SMEs. Individual-specific effects of SMEs is time-invariant and considered as an element of intercept. M. Ma (2016) indicated that individual-specific effect is constant under FE. So far, it is concluded that unobserved individual-specific effects of SMEs remain stable regardless of the shift in macroeconomics environment.

However, this research concentrates on the SMEs' change during the credit crisis and FE should be used whenever only analysing the impact of variables that vary over time and ignoring the unobserved effects. Additionally, one side effect of the features of FE models is that they cannot be used to investigate time-invariant causes of the dependent variables. Technically, time-invariant characteristics of the individuals are perfectly collinear with the individual dummies. Substantively, FE models are designed to study the causes of changes within an individual. A time-invariant characteristic cannot cause such a change, because it is constant for each individual (Kohler and Kreuter, 2005). Moreover, a constant individual-

specific effects of SMEs is impractical since the unobserved effects would be the relationship with banks, suppliers and others that could have impact on the defaulted event, and this event would change significantly during the credit crisis (M. Ma, 2016). Therefore, FE would not be a suitable solution to problems of heterogeneity.

On the other hand, the differences among individual-specific effects of SMEs lies in their individual specific errors instead of their intercepts. RE should be used when it is believed that differences across entities impact on dependent variable. Hence, RE would be preferred in this research.

3.7.2 Macroeconomic Variables (MVs)

To capture the time series effect, adding dummy variables would be an option (N. Beck, Katz, & Tucker, 1998). Yet, it is difficult to explain dummy variables and cannot apply outside the observed period. Macroeconomic variables (MVs) is another option to mimic the market movement during the credit crisis, and it is available to provide information of how MVs influence SMEs' performance. A key issues of accurately prediction is to link PD and MVs so that a panel regression is able to perform to capture a wide range of details not available when granular information is missing (Bellini, 2019).

To avoid additional and unnecessary influence brought from MVs, and control for market movement only, significant MVs will be added in for analysis. According to (Figlewski, Frydman, & Liang, 2012), they summarized the literature review that supports the view that credit risk exposure is influenced by conditions in the macroeconomic. Besides, they grouped potential MVs into three broad classes: general macroeconomic conditions, direction of the economy moving, and financial market conditions.

3.8 Model Performance

The receiver operating characteristics (ROC) curve , first proposed in the 1950s, is a widely used tool for evaluating discriminative and diagnostic power (Long, Zhang, & Hsu, 2011). ROC plots provide a threshold independent method of evaluating the performance of good/bad models. In a ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate (100-Specifiity) for different cut-off point. Each point on the ROC curve represents a sensitivity/specificity pair regarding a specific cut-off benchmark. The area under the ROC curve (AUROC) is a measure of how well a parameter can distinguish between two groups (good/bad).

When one uses the ROC curve to measure the performance of default prediction, the threshold is every cut-off probability, the ROC curve defines the “true positive rate” (percentage of defaults that the model correctly predicts as default) on the y-axis as a function of the corresponding “false positive rate” (percentage of non-defaults that are mistakenly predicted as default or other exits) on the x-axis.

First, in order to construct the ROC curve, all firms are ordered by their default probabilities from highest to lowest. Then at each default probability γ we calculate a set of two fractions, the first one is the percentage of defaults that the model correctly predicts as default, where the correct prediction means the default probability of the firm is equal or greater than γ.

$$f^{y}(\gamma) = \frac{\text{number of correctly predicted as default at the threshold } \gamma}{\text{number of total correctly predict default}} \qquad (42)$$

The second one is the percentage of non-defaults that are mistakenly predicted as default, also here the firm has a default probability that equal or greater than γ.

$$f^{x}(\gamma) = \frac{\text{number of mistakenly predicted as default at the threshold } \gamma}{\text{number of total mistakenly predict default}} \qquad (43)$$

As shown in Figure 3-3, the best possible test (100% sensitivity and 100% specificity) would have an area under curve of 1 (100%), but it is always too good to be true, and it may lead to overfitting. ROC analysis provides ways to choose possibly optimal models and to discard suboptimal models independently from the

cost context or the class distribution. ROC analysis is related in a direct and natural way to analysis of diagnostic decision making.

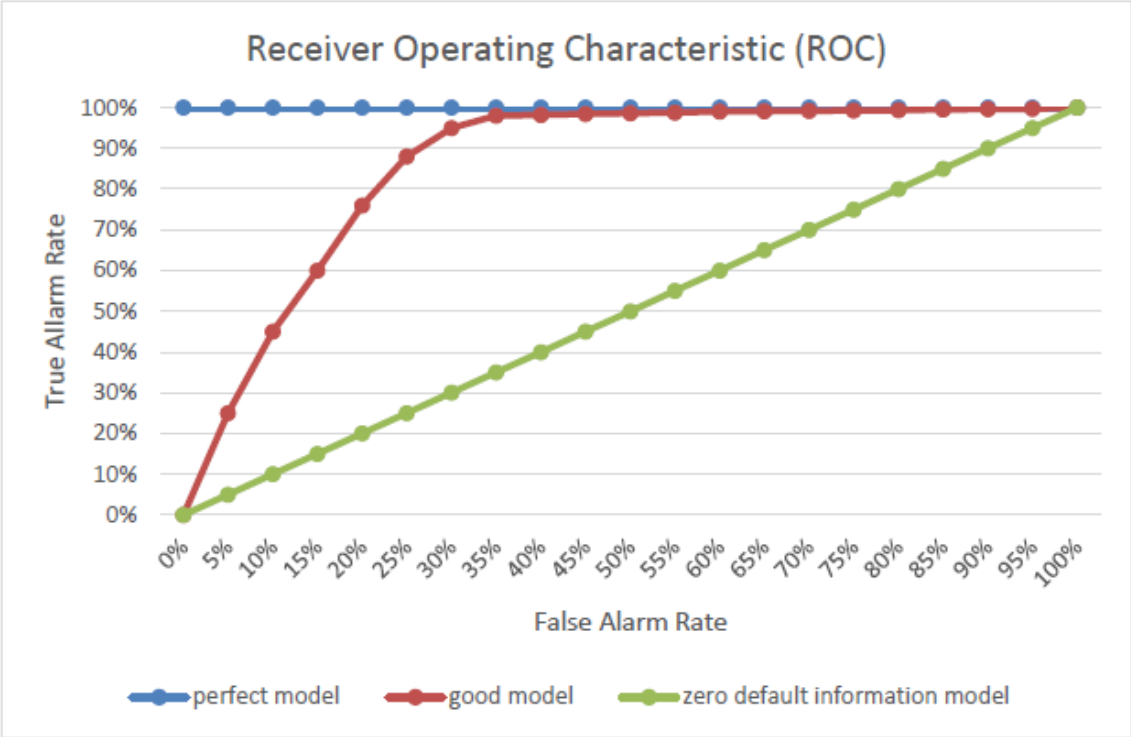


Figure 3-3 ROC of estimated default probability

Notes: Figure 3-3 plots the receiver operating characteristic of the estimated default probability. If the model contains no default information, then the ROC curve (the green line) will be corresponds to the 45-degree line, however a perfect model (the blue line) will have a ROC curve that goes straight up from (0, 0) to (0, 1) and then across to (1, 1). And a good model will be more like the red line.

3.9 Summary

This chapter has presented the methodological procedure starting with handling incomplete observations. MICE is a competitive and flexible method that can deal with missing variables individually. Afterwards, the credit risk models for SMEs are built up according to the cross-section analysis with the theoretical foundation of Logistic regression, Shrinkage regression models and GAMs, and there is a comparison between the WoE and stacked imputed data. Finally, the panel data

models include macroeconomic variables selection, the use of macroeconomic variables and the impact of the change in macro environment during the credit crisis on SMEs viability is discussed in detail.

In the following, Chapter 4 provides a description of the SMEs dataset used in this thesis. The first section in Chapter 5 presents the results of dealing with missing data using MICE and determines the predictors for credit risk modelling. The second chapter shows the finding of cross-section analysis and the third section is the results considering the change in macroeconomic conditions. A discussion of all findings is illustrated in the last section.

CHAPTER 4 DATA DESCRIPTION

4.1 Introduction

The methodology has been introduced in the last chapter. Solving the missing data problem and credit risk modelling are essential to this research. After the credit crisis, the Basel Committee's Principle for Risk Data Aggregation and IFRS 9 pose strong governance on data usage and reporting. This chapter presents a comprehensive description of the data that regard as a source of predictors about SMEs performance. By further dividing SMEs as start-ups and non-start-ups, their primary difference is observed.

4.2 Sample data

The dataset collects UK SMEs' information between 2007 and 2010 with over 80 different characteristics, including general information, directors' information, previous relevant credit history and accounting information. Besides, the observed period is classified as three blocks: 2007 regular economic period (pre-crisis period), 2008-2009 financial crises period, 2010 recovery period (post-crisis period).

Dietsch and Petey (2004) indicated that it is reasonable to distinguish different segments inside the SME's population. Orton, et al. (2017) further classified SMEs into start-ups and non-start-ups for a separate analysis. Start-ups are established less than or equal to 60 months, those greater than 60 months are considered as non-start-ups. In general, non-start-ups SMEs would be more developed and mature, but newly established firms struggle more at their beginning (Baum, Calabrese, & Silverman, 2000) due to a lack of stable relationships and sufficient resources. Yet, (Hudson, 1987) pointed out that a newly formed company is most likely to have a "honeymoon period" of around 2 years before being in real risk. Hence, it is acceptable to analysis these two segments separately assuming that they have a remarkable difference. Characteristic comparisons, such as the 1992

SIC code, and regions, used to identify the difference between start-ups and non-start-ups are provided in the later section.

For some characteristics, unreasonable observations are easy to detect. For example, the total fixed assets as a percentage of total assets should be within a range from zero to hundred. Hence, modification is necessary if observations are out of the range. If the ratio is greater than 100, then set as 100, and if its value is smaller than 0, then set as 0. Finally, the data selection process is summarised as followed:

- Delete the observations that violate SMEs definition by the European Commission
- Adjust the scale of specific variables
- Fifty per cent of SMEs were randomly select and ensure that the ratios of good/ bad are equivalent to the population.

4.3 “Bad” Rates

In most credit scoring models the dependent variable is binary: the borrower is either non-default or default. In the given dataset, those SMEs with “bad” flag does not necessarily mean defaulted, yet their credits are heavily impaired. As in the three-stage model in IFRS 9, it is likely that those SMEs with “bad” flag fall into stag 2 even stage 3, and a corresponding lifetime ECL allowance is recognised. Banks should be worried about the ability of those SMEs to repay the loan as it has shown signs of significant deterioration in credit quality. The status of SMEs in the dataset has been labelled as either “good” or “bad” and this research focuses on binary classification problem, for simplicity, SMEs with “bad” flag are intuitively regarded as default, while those with “good” flag are considered as non-default.

The last credit crisis had a severe influence on SMEs resulted in a sharp rise of ‘bad’ rate from 2007 to 2009, see Table 4-1. Regarding the total number of SMEs in the sample dataset, there were roughly balanced observations during the observed period between the start-ups and the non-start-ups. Initially, the total number of start-ups SMEs exceeded that of non-start-ups. After that, an increasing

trend was observed for non-start-ups, while a decreasing trend for start-ups and the turning point that the number of start-ups SMEs surpassed that of non-start-ups was observed at the peak of the credit crisis. In the post-crisis period, the number still rose only for non-start-ups. This situation may indicate that non-start-ups is more capable of survival in the last credit crisis.

Table 4-1 Frequency of UK SMEs data

Year	Start-ups				Non-start-ups			
	Good	Bad	Total	Bad(%)	Good	Bad	Total	Bad(%)
2007	42758	4552	47310	9.62	38991	2013	41004	4.91
2008	39412	7911	47323	16.72	42009	4210	46219	9.11
2009	34402	8951	43353	20.65	41753	7352	49105	14.97
2010	33480	6960	40440	17.21	44575	5007	49582	10.10

Notes: The first column shows the year
The next four columns list the details of start-ups sample
The last four columns list the details of non-start-ups sample.

The “bad” rate (percentage of the defaulted firm) of start-ups was always higher than that of non-start-ups during the observed period. The default rate for non-start-ups began at less than 5% and then peaked at 14.97% in 2009 before slightly falling to about 10% in the recovery period. Likewise, start-ups saw around 9.62% default rate in 2007 following a significant rise to about 20% in 2009 and falling to 17.21% in the last year. In summary, although ‘bad’ rate of both start-ups and non-start-ups experienced a similar trend, ‘bad’ rate of non-start-ups was considerably lower than that of start-ups for each year, and they both reached an all-time peak in 2009, then dropped down to 2008 level.

Overall, the financial crisis began in 2007 in the USA, and its adverse effects soon spread globally. In 2009, UK SMEs experienced the most severe impact leading a high number of SMEs being defaulted, after that the situation of economic depression began to recover since 2010. Data evidence has shown that the number of start-ups sharply decreased as they were vulnerable, and credit crisis had less impact on non-start-ups than that of start-ups. In summary, due to the large gap of the default rate in each period, this initially suggests that some factors determine their performance during the credit crisis.

4.4 Variables explanation

A distinct difference of the default rate between start-ups and non-start-ups is found in the previous section. In the following subsections, the characteristics leading the difference are displayed and discussed.

4.4.1 1992 SIC Code

Industry classification is one of the important features that lead to a diversity of SMEs performance. This dataset recorded the UK 1992 Standard Industrial Classification Codes (1992 SIC Codes) to divide SMEs into various industry categories[32]. This standard is the most widely used during the observed period.

Table 4-2 shows the frequency of start-ups and non-start-ups by industry sectors. The primary difference between these two segments is the percentage of missing data group (Group NA). Compared with non-start-ups, start-ups SMEs were reluctant to report their industry classification. In addition, Real Estate, Renting & Business Activities (Group 10) and Public Administration & Defence (Group 11) also showed a huge difference. In concluded, there were significant differences in industry distribution for start-ups and non-start-ups.

The “bad” rate across different industry sectors for non-start-ups and start-ups SMEs respectively from 2007 to 2010 is shown in Table 4-3. For both segments, the default rate of most industry sectors reached its highest point in 2009, and then slightly declined in 2010. The inelastic demand sectors remained a low default rate for both segments. Besides, the credit crisis should have a material shock on the financial industry, yet it is surprising to see that Financial Intermediation (Group 9) sector had a low ‘bad’ rate. Both Real Estate, Renting & Business Activities (Group 10) and Public Administration & Defence (Group 11) saw a relative high default rate. For start-ups, missing group (Group NA), occurred for over 50% observations,

[32] http://www.ukmarketingmanagement.com/mailing-lists/business-lists/sic-codes/

was worth to pay attention as its default rate far exceeded that of other sectors, especially in 2009 approximately 13.97% of default rate (total default rate in 2009 was 20.65%). For non-start-ups, Public Administration & Defence (Group 11) had a high default rate. Its default rate began at around 1% after that over 6% of that industry defaulted in 2009 before a dramatic decrease to 2.7 % in the following year. Besides, Real Estate, Renting & Business Activities (Group 10) suffered severely. The default rate of real estate, renting and business activities were 2.82% in 2009, in second place overall.

Overall in comparison of these two groups, there was a significant gap in the default rate of missing data group. Specifically, the default rate of the missing group for non-start-ups SMEs was much lower than that in start-ups SMEs. Also, Real Estate, Renting & Business Activities (Group 10) and Public Administration & Defence (Group 11) suffered the most negative impact for both for start-ups and non-start-ups SMEs.

Table 4-2 Percentage (%) table of industry sectors

Year	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	NA
Start-ups																
2007	0.20	0.03	0.11	1.50	0.03	3.69	3.50	1.24	1.56	0.62	9.73	14.70	0.54	0.87	3.30	58.39
2008	0.25	0.03	0.08	1.30	0.05	3.70	3.46	1.18	1.40	0.66	10.30	12.80	0.47	0.88	3.22	60.21
2009	0.27	0.03	0.10	1.37	0.07	4.24	3.42	1.30	1.43	0.67	10.80	12.83	0.51	1.07	3.37	58.53
2010	0.27	0.04	0.11	1.33	0.10	4.04	3.57	1.22	1.44	0.58	9.80	11.51	0.49	1.18	3.38	60.93
Non-start-ups																
2007	0.83	0.11	0.26	5.88	0.08	9.73	10.52	2.90	3.16	1.63	25.83	21.36	1.12	2.15	10.90	3.54
2008	0.77	0.10	0.22	5.41	0.09	9.65	10.50	2.89	3.07	1.63	25.51	22.91	1.19	2.05	10.42	3.59
2009	0.81	0.13	0.20	4.96	0.09	9.91	10.40	2.80	3.04	1.66	25.41	23.33	1.32	2.03	10.29	3.64
2010	0.86	0.12	0.24	4.75	0.13	10.48	10.20	2.69	2.96	1.61	26.23	22.04	1.34	2.39	9.77	4.22

Notes: Frequency across different 1992 SIC code. The first column indicates the year, and the first row represents different industry sectors, and its meaning is given below. The upper panel presents the percentage of SMEs in different industry sectors for start-ups, while the lower panel presents that for non-start-ups.

0: Agriculture, Hunting and Forestry, 1: Fishing, Mining & Quarrying, 2: Manufacturing, 3: Electricity, 4: Gas & Water Supply, 5: Construction, 6: Wholesale & Retail Trade, 7: Hotels & Restaurants, 8: Transport, Storage & Communication, 9: Financial Intermediation, 10: Real Estate, Renting & Business Activities, 11: Public Administration & Defence, 12: Education, 13: Health & Social Work, 14: Other Community, Social & Personal Service Activities, NA: Missing

Table 4-3 “bad” rate (%) across different industries

Year	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	NA
Start-ups																
2007	0.00	0.00	0.01	0.09	0.00	0.21	0.25	0.10	0.09	0.03	0.50	0.86	0.05	0.11	0.15	7.16
2008	0.02	0.01	0.01	0.14	0.01	0.42	0.40	0.18	0.16	0.06	1.29	1.68	0.03	0.11	0.31	11.89
2009	0.03	0.00	0.01	0.34	0.01	0.50	0.48	0.25	0.22	0.08	1.42	2.51	0.06	0.15	0.61	13.97
2010	0.02	0.00	0.01	0.24	0.01	0.60	0.52	0.12	0.17	0.07	1.21	1.98	0.06	0.19	0.45	11.56
Non-start-ups																
2007	0.04	0.00	0.01	0.31	0.00	0.49	0.59	0.19	0.17	0.06	1.26	1.09	0.05	0.08	0.40	0.16
2008	0.04	0.00	0.03	0.36	0.01	0.88	0.98	0.36	0.32	0.11	2.44	2.40	0.10	0.13	0.75	0.20
2009	0.08	0.01	0.02	0.52	0.01	1.17	1.29	0.51	0.66	0.22	2.82	6.25	0.11	0.18	0.89	0.25
2010	0.04	0.00	0.02	0.46	0.01	1.05	0.98	0.41	0.33	0.15	2.38	2.77	0.10	0.17	0.69	0.53

Notes: Default rate (%) across different 1992 SIC code. The first column indicates the year, and the first row represents different industry sectors, and its meaning is given below. The upper panel presents the ‘bad’ rate of SMEs in different industry sectors for start-ups, while the lower panel presents that for non-start-ups.

0: Agriculture, Hunting and Forestry, 1: Fishing, Mining & Quarrying, 2: Manufacturing, 3: Electricity, 4: Gas & Water Supply, 5: Construction, 6: Wholesale & Retail Trade, 7: Hotels & Restaurants, 8: Transport, Storage & Communication, 9: Financial Intermediation, 10: Real Estate, Renting & Business Activities, 11: Public Administration & Defence, 12: Education, 13: Health & Social Work, 14: Other Community, Social & Personal Service Activities, NA: Missing

4.4.2 Regions

There is a different effect due to SMEs' locations. Changeable regional policy, economic environment and business conditions have an impact on the development of SMEs, especially during the credit crisis. Federico, Rabetino, & Kantis (2012) showed that regional difference influence SMEs' evolution. For example, London may be a suitable choice of wholesale and retail trade given its unique financial status, but it may not a reasonable option for manufacturing industry due to its limited space and expensive land cost.

Twelve regions across the UK are shown in Table 4-4, and the other category (Group 13) refers to firms which could not be classified into any regions. There was a small proportion of missing data (Group NA). SMEs located in London (Group 1) and South East regions (Group 2) accounted for approximately 40% of SMEs for both two segments in each year.

The default rate across different regions is presented in Table 4-5. Likewise, the year 2009 saw the highest default in majority areas. It is no doubt that London must experience the enormous shock because of the scale of its financial industry. The default rate of London for non-start-ups was more stable than that of start-ups.

For non-start-ups, South East (Group 2) region as the second largest regional economy in the UK (after London), ranked the first place in terms of default rate in 2008 and 2009. There was an increase of default rate from 2008 to 2009, ranging from 1.92% to 4.89% in South East region; more than five times the default rate in comparison to that in the year 2007. The default rate changed sharply in the North West region as well, jumping from 0.51% in 2007 to 1.46% in 2008.

In summary, non-start-ups had a lower default rate. The South East region suffered most for non-start-ups while London area suffered most for start-ups SMEs.

Table 4-4 Percentage table of regions

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	NA
Start-ups														
2007	23.40	18.18	6.36	1.74	12.24	3.61	8.63	9.54	7.08	4.66	2.16	0.88	0.01	1.51
2008	24.30	16.02	6.68	1.83	12.00	3.91	8.56	10.36	6.57	5.09	2.38	0.87	0.02	1.40
2009	24.27	15.71	6.41	2.09	11.05	4.29	8.55	10.87	6.52	5.46	2.38	1.00	0.01	1.39
2010	24.76	16.18	6.67	2.15	10.03	3.88	8.01	11.01	6.58	5.69	2.61	1.00	0.01	1.41
Non-start-ups														
2007	17.99	19.83	8.23	2.11	9.46	4.50	7.21	11.41	6.81	5.87	2.54	1.23	0.03	2.78
2008	17.24	20.38	8.21	2.22	9.49	4.59	7.41	11.31	6.77	5.96	2.55	1.33	0.02	2.52
2009	17.14	20.09	8.18	2.16	9.62	4.51	7.82	11.37	6.83	6.08	2.52	1.29	0.03	2.36
2010	17.78	17.35	8.47	2.28	10.53	4.58	8.27	11.74	6.50	6.18	2.76	1.50	0.03	2.03

Notes: Percentage (%) across regions between 2007 and 2010. The first column indicates the year, and the first row represents different regions, and its meaning is given below. The upper panel presents the percentage of SMEs in different regions for start-ups, while the lower panel presents that for non-start-ups.

1: London; 2: South East; 3: South West; 4: North East; 5: North West; 6: East Midlands; 7: West Midlands; 8: East England; 9: Yorkshire; 10: Scotland; 11: Wales; 12: North Ireland; 13: Others; NA: Missing

Table 4-5 ‘bad’ rate across different regions

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	NA
Start-ups														
2007	2.90	1.78	0.43	0.15	1.02	0.27	0.77	0.89	0.75	0.35	0.17	0.02	0.00	0.12
2008	4.72	2.81	0.78	0.22	2.72	0.51	1.18	1.52	1.36	0.39	0.31	0.05	0.00	0.16
2009	6.18	2.95	1.05	0.40	2.38	0.78	1.90	2.05	1.33	0.94	0.42	0.06	0.00	0.22
2010	4.85	2.50	0.95	0.33	1.78	0.56	1.49	1.77	1.23	0.88	0.46	0.18	0.00	0.21
Non-start-ups														
2007	0.98	0.86	0.37	0.11	0.51	0.19	0.40	0.57	0.32	0.29	0.13	0.05	0.00	0.13
2008	1.91	1.92	0.62	0.20	0.86	0.37	0.75	1.02	0.66	0.36	0.23	0.06	0.00	0.15
2009	2.25	4.89	0.84	0.30	1.46	0.60	1.19	1.33	1.01	0.58	0.29	0.04	0.00	0.19
2010	1.93	1.58	0.80	0.24	1.04	0.43	1.18	1.09	0.58	0.63	0.24	0.24	0.00	0.13

Notes: Default rate (%) across regions between 2007 and 2010. The first column indicates the year, and the first row represents different regions, and its meaning is given below. The upper panel presents the ‘bad’ rate of SMEs in different regions for start-ups, while the lower panel presents that for non-start-ups.

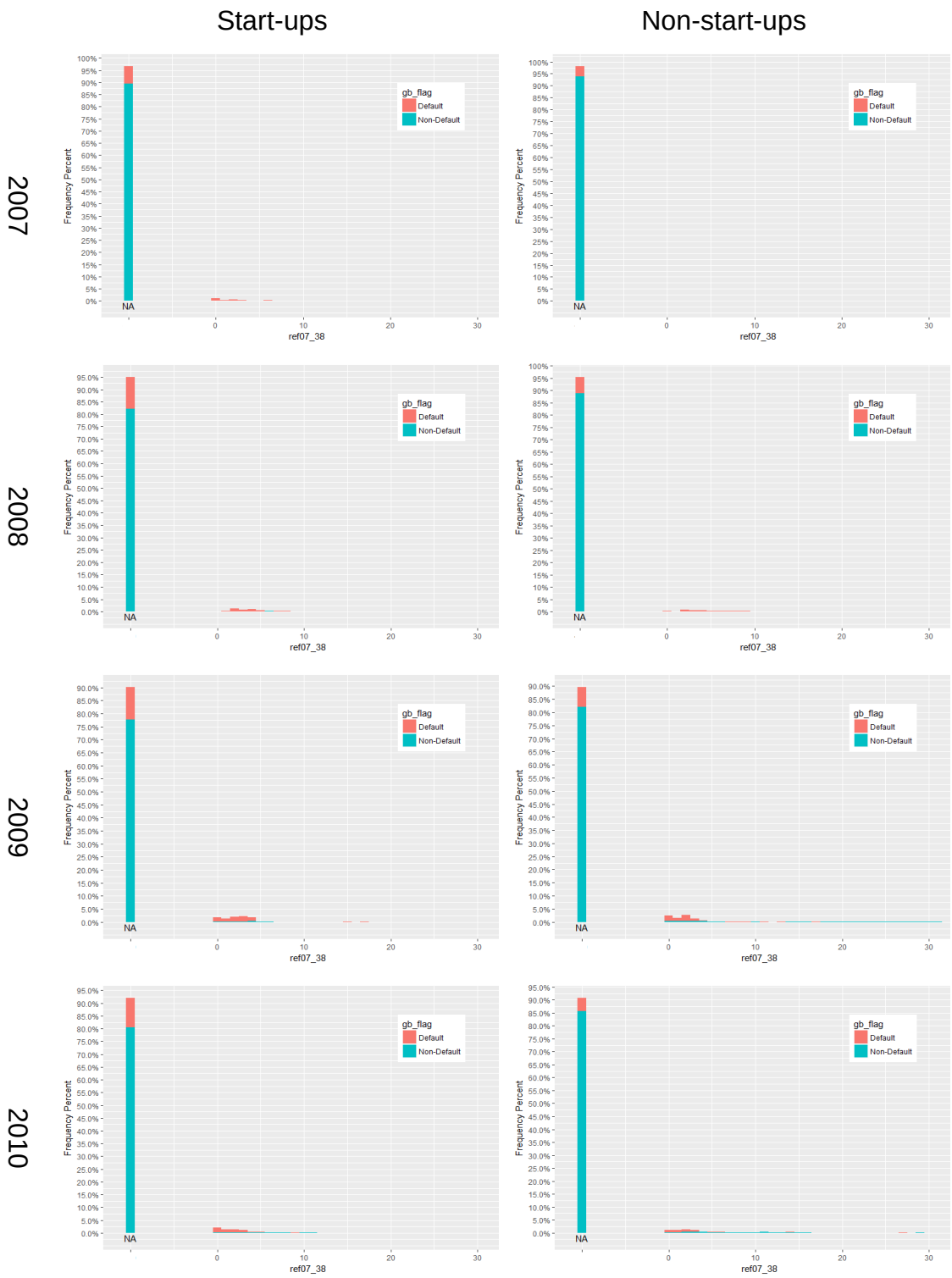
1: London; 2: South East; 3: South West; 4: North East; 5: North West; 6: East Midlands; 7: West Midlands; 8: East England; 9: Yorkshire; 10: Scotland; 11: Wales; 12: North Ireland; 13: Others; NA: Missing

4.4.3 Time since Last Derogatory Data Item (Months)

It is necessary to explore the sensitive variable time since last derogatory data item (months) since it is the variable with the largest missing proportion for both start-ups and non-start-ups. A derogatory item is negative and typically indicates serious delinquency or late payments. Derogatory items represent credit risk to lenders, and therefore, are likely to have a substantial effect on the ability to obtain new credit for borrowers. Public record items, such as bankruptcies, tax, and judgments also are considered derogatory. While some lenders still may be willing to extend credit to someone with derogatory items on their report, they may do so with higher interest rates or fees. Therefore, it is intuitively expected that the shorter time since the last derogatory, the worst the credit quality is.

As shown in Figure 4-1, majority of observations were empty (Group NA, over 90%). Within missing group, most observations were classified as non-defaulted for both segments. Yet, the proportion of default in the missing group for start-ups was higher than that of non-start-ups.

Missing groups were removed as it takes up the majority of observations, see Figure 4-2. With the time goes by, the proportion of non-default was greater than the proportion of default for both start-ups and non-start-ups. It is not surprising that non-default attribute showed a “right skew” trend, which indicated that the longer time since the last derogatory, the better credit situation was. Excluding non-start-ups in 2007, the pattern of all periods remained stable. There was a great number of non-default of non-start-ups in 2007, yet the number of non-default and default sharply decreased during the financial crisis.



Notes: the most left bar indicates NAs (missing category)

Figure 4-1 Frequency percentage plots of time since last derogatory data item (months)

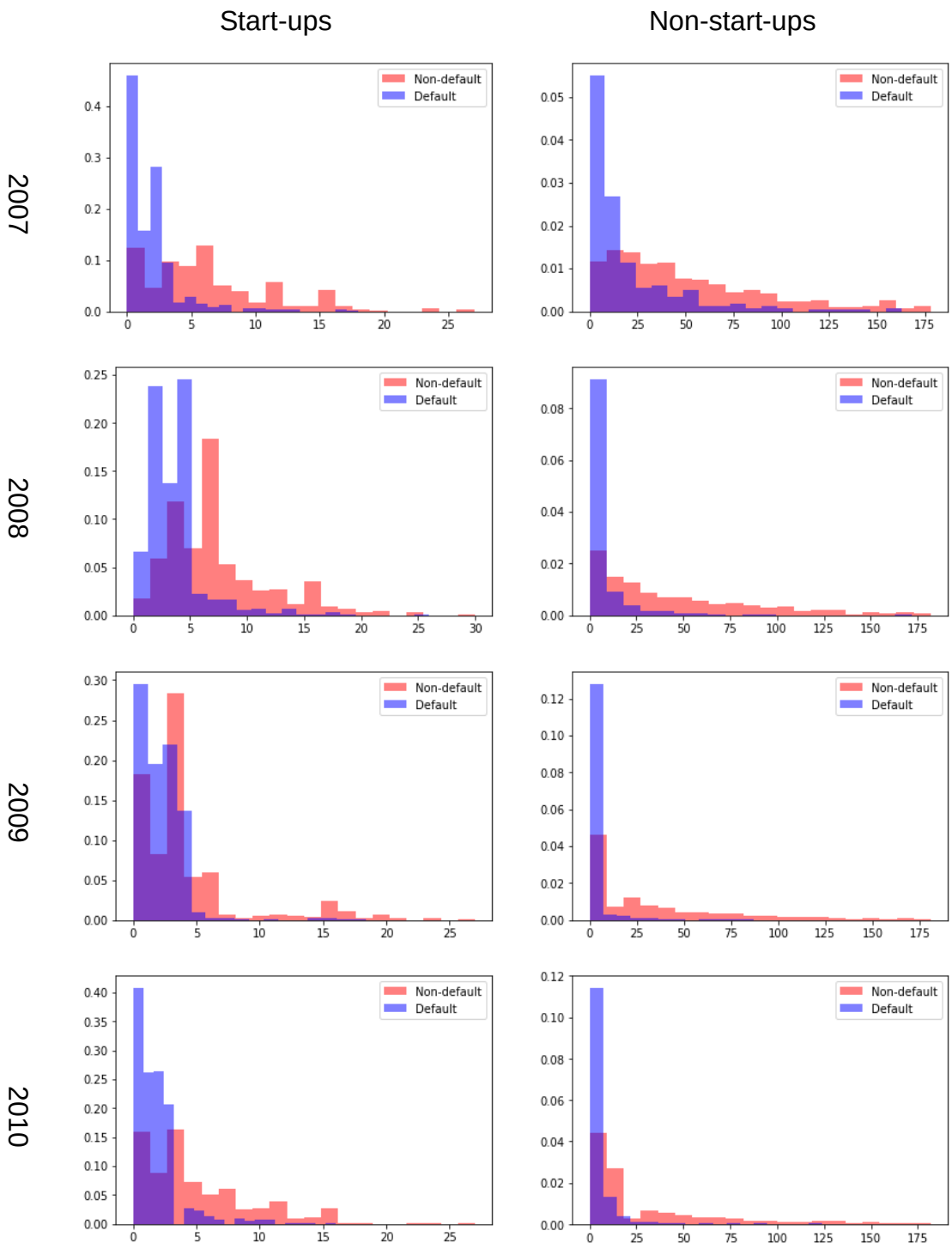


Figure 4-2 Frequency percentage plots of time since last derogatory data item (months) after removing the missing group

4.4.4 Proportion of Current Directors to Previous Directors in the Last Year

With the proportion increase, the number of new directors rises. Frequency percentage plots are presented in Figure 4-3. Both start-up and non-start-up were not willing to report the directors' information since a large volume of missing data (at least 85%) was observed. Within the missing data group, the default probability increased as the peak of the financial crisis approached (an increase of red area) and reached its top in 2009 for both start-ups and non-start-ups. Besides, the default probability for start-ups was slightly larger than that of non-start-ups.

For the non-missing group, start-up tended to be more conservative in changing the directors as a large number of observations were centred in the leftmost interval (excluding the missing category). Besides, the height of the leftmost interval decreasing with time from 2007 to 2009 indicated that start-ups preferred to remain stable management while non-start-ups were inclined to appoint new directors.

4.4.5 Time since Last Annual Return

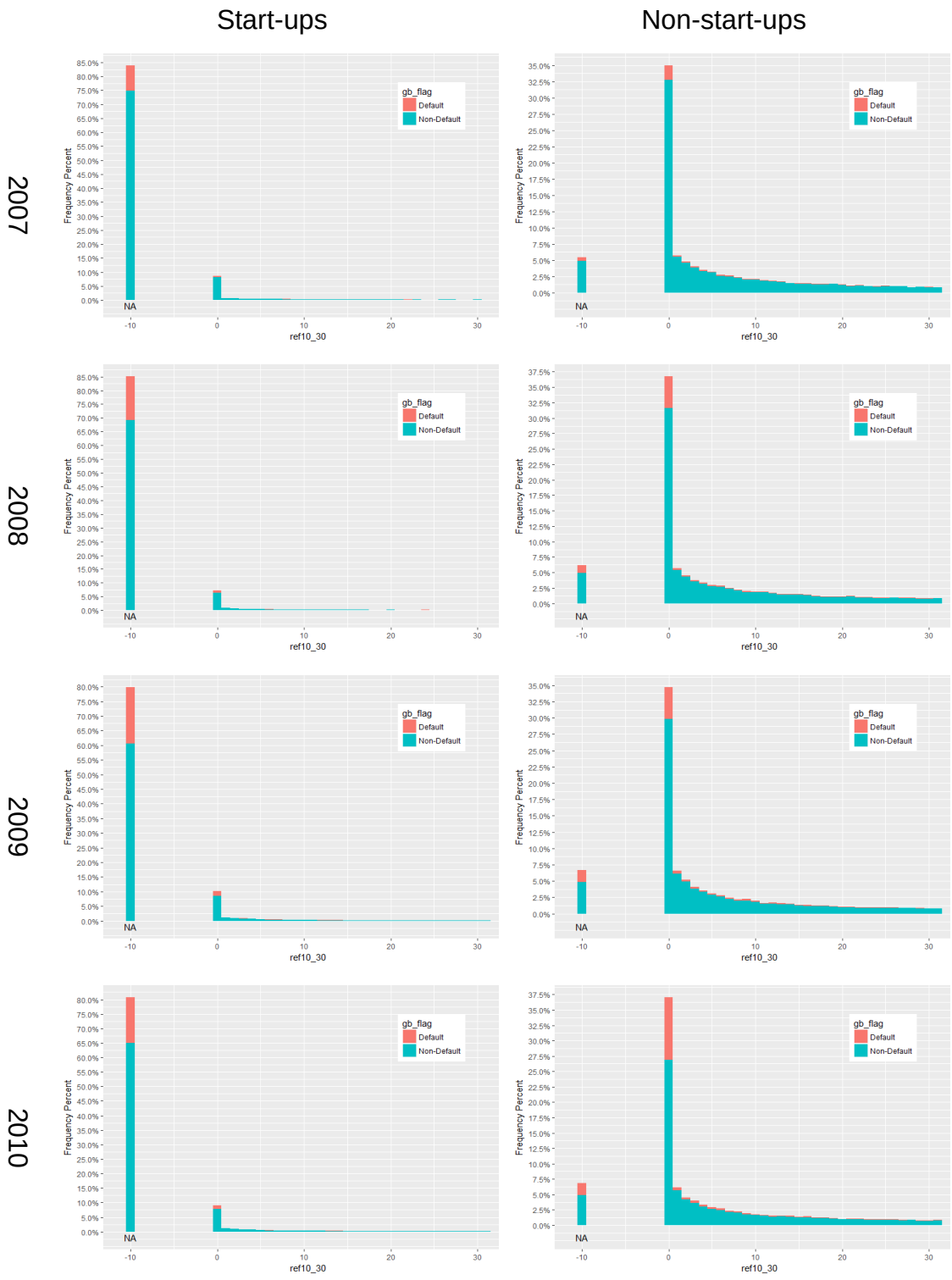
Instead of a financial document, an annual return (Confirmation statement) is a record of publicly available information about a firm that appears on the Companies Register. That information, which includes address and details of directors and shareholders, must be updated each year through Companies House[33].

A distinct difference in missing data category is found on Figure 4-4. For non-start-ups, the majority of observations was close to zero, which meant matured firms reported their annual return more frequently, while this did not hold for start-ups due to a large proportion of missing data.

[33] Running a limited company, https://www.gov.uk/running-a-limited-company/company-annual-return

Notes: the most left bar indicates NAs (missing category)

Figure 4-3 Frequency percentage plots of Proportion of current directors to previous directors in the last year



Notes: the most left bar indicates NAs (missing category)

Figure 4-4 Frequency percentage plots of time since last annual return

4.5 Summary

This chapter presents SMEs performance from 2007 to 2010. The ‘bad’ rate increased since the outbreak of the credit crisis and reached the peak at 2009, but the ‘bad’ rate of start-ups is much higher than that of non-start-ups. The ‘bad’ rate of the missing group for non-start-ups SMEs was much lower than that in start-ups SMEs. Also, Real Estate, Renting & Business Activities (Group 10) and Public Administration & Defence (Group 11) suffered the most negative impact for both for start-ups and non-start-ups SMEs. The South East region suffered most for non-start-ups while London area suffered most for start-ups SMEs. The distinct difference of default rate, industry performance, and regional performance and other factors shows that it is reasonable to further split SMEs as start-ups SMEs and non-start-ups SMEs to build up the credit risk modeling.

CHAPTER 5 RESULTS

5.1 Imputation

The problem of missing data is addressed by the 'state of the art' MICE technique under the assumption that the missing mechanism is ignorable. This can ensure that the imputation model preserves the relationships among the variables of interest (Moons, Donders, Stijnen, & Harrell, 2006), and variables in the analysis model are also included in the imputation model, thus leading to unbiased estimates (Angela M Wood, et al., 2008).

This research employs MICE package in R, and the seed is set to be a fixed number so that the results are reproducible. As previously mentioned, it is necessary to consider the computation time of MICE as the MICE procedure would be time-consuming if using the whole population. Therefore, random subset is selected for keeping balance in reasonable computation time and availability of imputation results for each year, and it takes one minute on average for each iteration.

5.1.1 Results of MICE Imputation

MICE generates 50 imputed datasets with 20 maximum iterations in which continuous variables, binary variables and categorical variables are imputed by PMM, logistic regression, and multinomial logistic regression respectively. Although some of the continuous variables are extremely skewed or semi-continuous, they are imputed on the raw scale (i.e., without transformation) regardless of their distribution (Von Hippel, 2013; Katherine J Lee and Carlin, 2017).

5.1.1.1 Pooling Results

Table 5-1 provides coefficients, standard errors and fraction of missing information of pooled variables coefficients after MICE imputation, repeatedly running logistic regression until all variables are significant. Totally twenty-three variables are selected as predictors into the model, of which nine, twelve, sixteen, seventeen variables are picked from 2007 to 2010, respectively.

Fifteen of total twenty-three variables are selected once, and eight variables are significant for more than two years where only five variables are significant four years in a row. These five variables are time since last derogatory data item (months), time since last annual return, total fixed assets as a percentage of total assets, the second category of incomplete directors information flag, and the fifth category of last derogatory item.

Time since the last derogatory data item (months) is significant at 1% level over the four-year period. A derogatory item is considered to be a negative variable, and typically indicates serious delinquency or late payments. Derogatory items represent credit risk to lenders, and therefore, are likely to have a massive impact on the ability to grant new credit. Intuitively, the longer the time is, it is more likely to be non-defaulted. Thus, it is not surprising that this variable has a reverse relationship to default (positive coefficient) and become an essential predictor after imputation in any observed period. However, given it has a much higher FMI, increasing numbers of imputation would obtain a fairer and more convincing estimate.

Time since last annual return is significant at the 10% level in 2007, and it becomes significant at the 1% level from 2008 to 2010. This variable illustrates the time since the last report the firm performance to Companies House, and it is said that a healthy SME would report performance in time and more frequently. The coefficient in 2009 is -4.22, which is much larger than that in other years. During the peak of the crisis, reporting performance had the most considerable on the SMEs' performance.

Likewise, total fixed assets as a percentage of total assets is significant at 10% level in 2007, after that its significant level increased gradually. A fixed asset is defined as an asset, which is a long-term tangible piece of property that a firm owns and uses in its operations to generate income. Fixed assets are not expected to be consumed or converted into cash within a year. Thus, this variable shows the ability of assets to be liquid or not. SMEs with a large proportion of fixed asset means that it cannot immediately raise cash as their assets are more illiquid. The positive coefficient of all period means log odds of being good increases with the increase of this ratio. Cho, Chung, & Kim (2014) found that during the 2008–2010 global financial crisis, the Korean government allowed firms to revalue their fixed assets to strengthen their balance sheets, helping distressed firms to obtain external financing. Fixed asset revaluation is an effective policy tool in Korea for helping firms obtain long-term debt financing, and the benefits are greatly pronounced in firms with financial constraints. Therefore, during the credit crisis, fixed assets helped SMEs to survive.

The second category of incomplete directors' information flag is significant at 1% level during the whole period. Being allocated to the second category, versus the base category, changes the log odds of being good in a negative direction.

Similarly, it is likely to hurt one's ability to qualify for credit if derogatory item is found on the credit report. The negative coefficients indicate that the fifth category of derogatory item has a negative relationship with being "good" and is significant at 1% between 2007 and 2010.

Table 5-1 Pooled logistic regression results (the result of step 3 of MICE)

Variables	2007			2008			2009			2010		
	est	se	fmi	est	se	fmi	est	se	fmi	est	se	fmi
(Intercept)	7.64	1.32	0.77	5.05	0.65	0.67	8.39	0.9	0.73	7.55	1.22	0.58
Age of Company				-1.05	0.12	0.53	-0.57	0.22	0.76	-0.39	0.09	0.36
No. Of 'Current' Directors				0.61	0.09	0.35				0.55	0.07	0.18
Number of Appointments In The last 12 Months as a Percentage of the Current Board				-0.26	0.13	0.88				0.37	0.06	0.66
Highest number Of Current Other Directorships of The Current Board/Proprietors Supplied				-0.67	0.23	0.51						
Number of Directors Holding Shares							0.31	0.05	0.29			
Number of Previous Searches (last 3m)										0.34	0.1	0.12
Number of Previous Searches (last 6m)										0.22	0.11	0.15
Time since last derogatory data item (months)	1.05	0.23	0.83	1.82	0.18	0.71	1.05	0.35	0.96	1.7	0.25	0.8
Lateness of Accounts	-0.95	0.49	0.75									
No Days between Accounting Date of Latest Filed Accounts and Date Recorded At Companies House							0.19	0.07	0.83			
Number of Years Accounts Available							0.49	0.11	0.45			
Time since Last Annual Return	-0.78	0.43	0.67	-1.27	0.2	0.77	-4.22	0.39	0.78	-1.04	0.22	0.76
Total Fixed Assets as a Percentage of Total Assets	0.22	0.13	0.93	0.21	0.1	0.91	0.26	0.06	0.78	0.27	0.07	0.82
Base Trend of Shareholders Funds	0.22	0.06	0.35				0.27	0.05	0.28			
Full CAIS Delphi score*							0.2	0.06	0.73	0.26	0.04	0.44
Legal Form_2							-4.06	1.09	0.51			
Legal Form_3							-3.79	0.39	0.45			
Legal Form_5							-3.97	0.59	0.38			
Legal Form_7							-2.36	0.47	0.36			
1992 SIC Code_2	-0.78	0.26	0.53									
1992 SIC Code_4	-0.46	0.2	0.55							-0.37	0.15	0.51
1992 SIC Code_5				-0.37	0.2	0.56						
1992 SIC Code_6	-0.49	0.25	0.51							-0.47	0.16	0.41
1992 SIC Code_7							-0.72	0.24	0.55			
1992 SIC Code_8	-0.39	0.19	0.57				-0.58	0.15	0.49	-0.64	0.14	0.5
Accounts Audited_2										-2.05	0.96	0.51
Accounts Audited_3										-2.24	0.95	0.51
Accounts Qualified_2				-1.48	0.64	0.83						
Accounts Qualified_3				-1.95	1.03	0.89						
Accounts Qualified_4							-1.14	0.62	0.7			
Incomplete Directors Information Flag_2	-1.79	0.18	0.61	-1.77	0.26	0.85	-1.13	0.2	0.71	-0.75	0.14	0.53
Payment Pattern_2										-0.61	0.36	0.73
Payment Pattern_5							-0.84	0.23	0.78			
Payment Pattern_6										-0.83	0.33	0.86
Last derogatory data item_2				2.18	0.91	0.75						
Last derogatory data item_3				2.09	0.79	0.66				-1.13	0.66	0.68
Last derogatory data item_5	-4.79	1.33	0.78	-2.28	0.4	0.55	-3.46	0.62	0.89	-2.61	0.56	0.72
Last derogatory data item_9				3.19	0.69	0.76						
Last derogatory data item_11				1.67	0.57	0.72						
Type of Accounts_1	0.92	0.37	0.22	-1.15	0.51	0.57	1.09	0.56	0.44			
Type of Accounts_2				-1.84	0.64	0.65	1.15	0.61	0.43			
Type of Accounts_8	0.64	0.23	0.52									
S2										-0.22	0.1	0.23

Notes: est: estimate of coefficient; se: standard error; fmi: fraction of missing information
Orange: significant at 1% level, Yellow: significant at 5% level, Red: significant at 10% level

In addition, the number of appointments in the last 12 months as a percentage of the current board is not significant in the regular economic period and the peak of the credit crisis but became significant in 2008 and 2010 at different levels. The larger the ratio, the more the new directors are recruited. New directors from different backgrounds can bring a number of benefits to boards, including a unique set of human capital resources (Kesner, 1988), new ideas and better communication (Milliken and Martins, 1996), debate (Pearce and Zahra, 1991; Fondas and Sassalos, 2000), and corporate governance processes (Singh, Terjesen, & Vinnicombe, 2008), which help complement the board's existing capacities. Therefore, the appropriate appointment of directors is key to withstand

the crisis because they tend to enhance corporate strategy and decision-making. For example, Bank of America Corp. and Citigroup Inc. appointed new directors with momentous working experience in banking or financial oversight and a deep understanding of regulatory issues in 2009 and these directors brought a material effect on the institution's performance during credit crises (Fernandes and Fich, 2009).

On the other hand, the larger the ratio, the larger the board size is. Shukeri, Shin, & Shaari (2012) suggested that there is a positive relation between board size and return on equity. Yasser, Mamun, & Rodrigs (2017) also indicated a positive relation between board size and performance. However, De Andres and Vallelado (2008) found that the relationship between bank performance and board size is not linear, but an inverted U-shaped. For every one-unit increase in one appointment, the log odds of being good decreased in 2008 yet increased in the recovery period. Yet appointing new directors during crisis period did not help SMEs to survival probably because they are unfamiliar with the operation process of the company but help recovery.

Age of company is significant since 2008 and a unit increases in age of company would reduce the log odds of being good during the financial crisis and recovery period. It is incomprehensible that negative coefficients are found on this variable. This finding is contrary to the previous view that young companies are more vulnerable during the credit crisis, but is partly consistent with Edward I Altman, Sabato, & Wilson (2008). They found that companies aging 3-9 years are more vulnerable to failure, and this may indicate that the relationship between the probability of being good and age of company is piecewise or nonlinear.

Legal form is a unique predictor as it is only significant in the period of financial crisis. This implies that SMEs with different legal forms had different performances when the economic environment changed sharply. With regard to accounts qualified and number of directors holding shares, the former is only significant during the beginning of credit risk period, while the latter is during the peak. Six categories of 1992 SIC code are included in the analysis model. Detailed industry

classification becomes less important as the majority are significant during the regular economic period instead of during the financial crisis period. The binary variable of start-ups or non-start-ups (S) and account audited is only significant in the recovery period, while lateness of accounts is solely significant at the regular economic period.

5.1.1.2 Checking the Imputation Model

Prior to imputation diagnostics, one point with examining imputation diagnostics is that differences between the observed and imputed values do not necessarily imply a problem (Stuart, Azur, Frangakis, & Leaf, 2009).

According to significant variables in Table 5-1, Table 5-2 tabulates summary statistics of the observed and imputed variables with missing data. Figure 5-1 below shows plots of convergence and distribution comparison of those variables with a large difference of the mean between observed and imputed data. There is no clear-cut method for determining whether the MICE algorithm has converged. However, by plotting parameters against the iteration number, the different streams should be freely intermingled with each other, without showing any definite trends if convergence is observed.

It is not surprising that imputed values of the continuous variables do not exceed the range of observed values because of the PMM method. Both observed and imputed variables have the same range of values since the corresponding minimum and maximum are identical, and the majority of observed and imputed continuous variables have a similar mean and standard deviation.

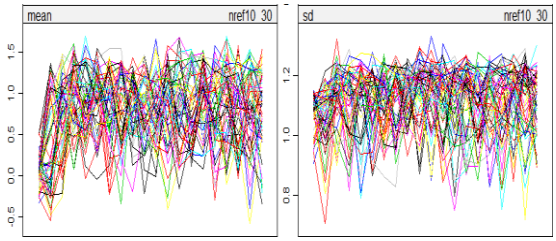
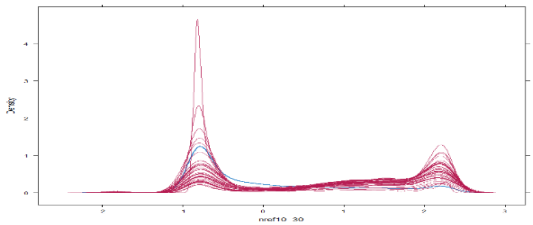
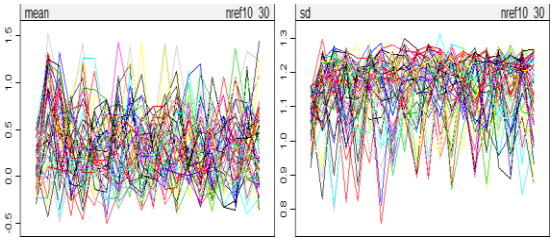
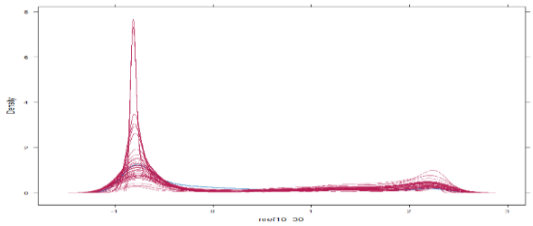
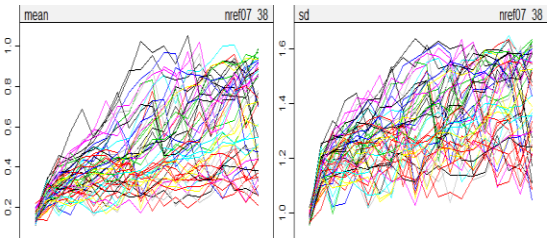
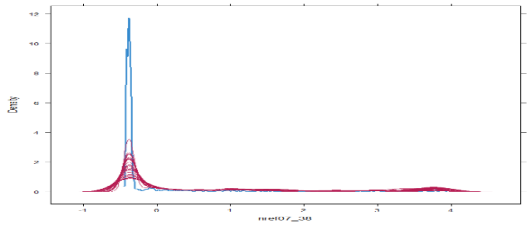
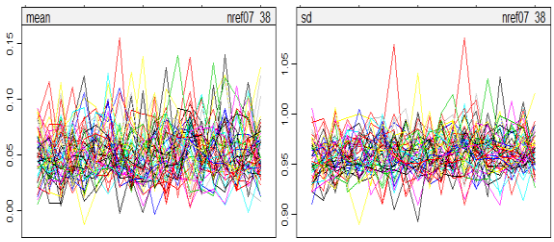
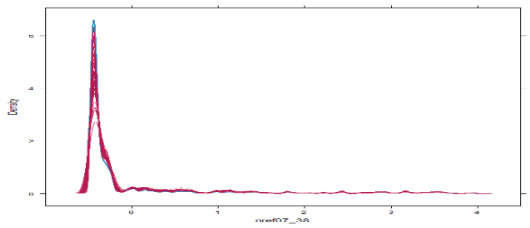
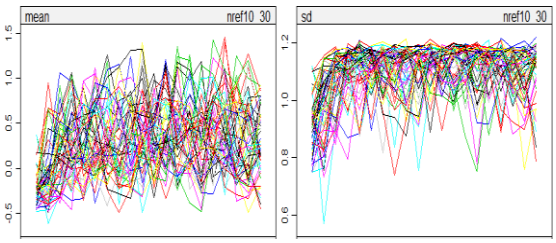
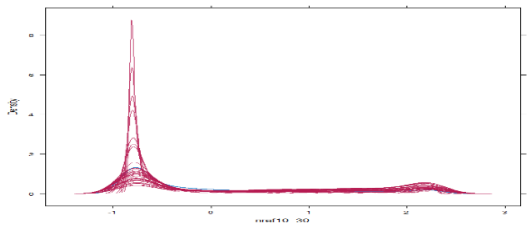
However, there are discrepancies between the observed and imputed values for a small number of variables. The variables with the largest difference of the mean (marked in red colour in Table 5-2) are further investigated by convergence plot and distribution comparison. In 2007 data, time since last derogatory data item (months) seems to be well imputed as its mean and standard deviation are close to the observed data, although it has the most substantial number of missing

values. The largest diversity of mean is found on the variable of total fixed assets as a percentage of total assets. As shown in Figure 5-1, the converge process of the mean and standard deviation are still not stable, which means they fails to converge. Similar kernel density plot of imputed values is observed because of the use of PMM. The same variable and situation occur in 2008 data. In 2009 data, the largest gap is found on time since last derogatory data item (months). Convergence plot has a strong initial trend and shows that the streams hardly mix and slowly resolve into a steady state. It is arduous and problematic to achieve convergence for a variable with high FMI and a large number of missing values. In 2010 data, the convergence plot of time since last derogatory data item (months) is not stable and that plot of total fixed assets as a percentage of total assets presents an increasing trend.

Table 5-2 Summary statistics of the observed and imputed data for the incomplete variables in the analysis model selected by Rubin's rules

Year	Variables	Observed					"Complete" (Imputed and Observed)			
		N	Mean	SD	Min	Max	Mean	SD	Min	Max
2007		17487								
	Time since last derogatory data item (months)	582	-0.1	0.9	-0.8	2.5	-0.3	0.7	-0.8	2.5
	Lateness of accounts	17148	-0.2	0.2	-0.4	4	-0.1	0.3	-0.4	4
	Time since last annual return	12066	-0.2	0.1	-0.3	4	-0.2	0.3	-0.3	4
	Total fixed assets as a percentage of total assets	10023	-0.03	1	-1.8	2.3	0.3	1.2	-1.8	2.3
2008		18190								
	No. Of 'current' directors	17822	-0.1	0.6	-0.4	3.9	-0.1	0.6	-0.4	3.9
	Number Of Appointments In The last 12 Months As A Percentage Of The Current Board	17822	0.03	1	-0.5	2	0.03	1	-0.5	2
	Highest number Of Current Other Directorships of the Current Board/Proprietors Supplied	17822	-0.1	0.2	-0.2	3.9	-0.1	0.2	-0.2	3.9
	Time since last derogatory data item (months)	986	-0.1	0.9	-0.6	3.2	-0.1	0.8	-0.6	3.2
	Time since last annual return	12850	-0.2	0.2	-0.3	3.9	-0.1	0.3	-0.3	3.9
	Total fixed assets as a percentage of total assets	10605	-0.02	1	-0.9	2.3	0.1	1.1	-0.9	2.3
2009		17858								
	Number of directors holding shares	17436	-0.04	0.9	-0.9	3.1	-0.1	0.9	-0.9	3.1
	Time since last derogatory data item (months)	1785	-0.1	0.8	-0.4	3.9	0.6	1.4	-0.4	3.9
	No. of Days between Accounting Date of Latest Filed Accounts and Date Recorded At Companies House	11463	-0.1	0.9	-2.3	4	-0.2	1	-2.3	4
	Time since last annual return	13313	-0.2	0.2	-0.3	3.2	-0.1	0.4	-0.3	3.2
	Total fixed assets as a percentage of total assets	11375	-0.1	1	-0.8	2.3	-0.01	1	-0.8	2.3
	Full CAIS Delphi score*	14187	0.1	0.8	-4	1.2	0.1	0.8	-4	1.2
2010		17784								
	No. Of 'current' directors	17415	-0.1	0.6	-0.5	3.8	-0.1	0.6	-0.5	3.8
	Number of Appointments In The last 12 Months as A Percentage of the Current Board	17415	0.03	1	-0.5	2	0.03	1	-0.5	2
	Time since last derogatory data item (months)	1545	-0.1	0.9	-0.5	4	0.04	1	-0.5	4
	Time since last annual return	13152	-0.2	0.1	-0.3	3.9	-0.2	0.3	-0.3	3.9
	Total fixed assets as a percentage of total assets	11640	-0.04	1	-0.8	2.3	0.1	1.1	-0.8	2.3
	Full CAIS Delphi score*	14426	0.1	0.8	-4	1.1	0.04	0.9	-4	1.1

Notes: The summary statistics of the "complete" data are calculated using pooled data over 50 imputations. SD standard deviation

Convergence plot (mean and standard deviation)	Distribution comparison plot
2007 Total fixed assets as a percentage of total assets	
	
2008 Total fixed assets as a percentage of total assets	
	
2009 Time since last derogatory data item (months)	
	
2010 Time since last derogatory data item (months) and Total fixed assets as a percentage of total assets	
	
	

Notes: Left panel: convergence plot of variables' mean and standard deviation. The total iteration is 20. Right panel: graphs comparing the distribution of the observed and imputed values. The blue line (observed values) and the red line (imputed values)

Figure 5-1 Plots of convergence and distribution comparison

Notes:

Imputed bar char is plotted by using pooled data over 50 imputations

The proportion of default (bad) versus non-default (good) is shown in each bar by colour.

A: bar chart of observed values keeping Missing values

B: bar chart of observed values removing Missing values

C: bar chart of imputed values

Figure 5-2 Bar chart of observed and imputed values of last derogatory item

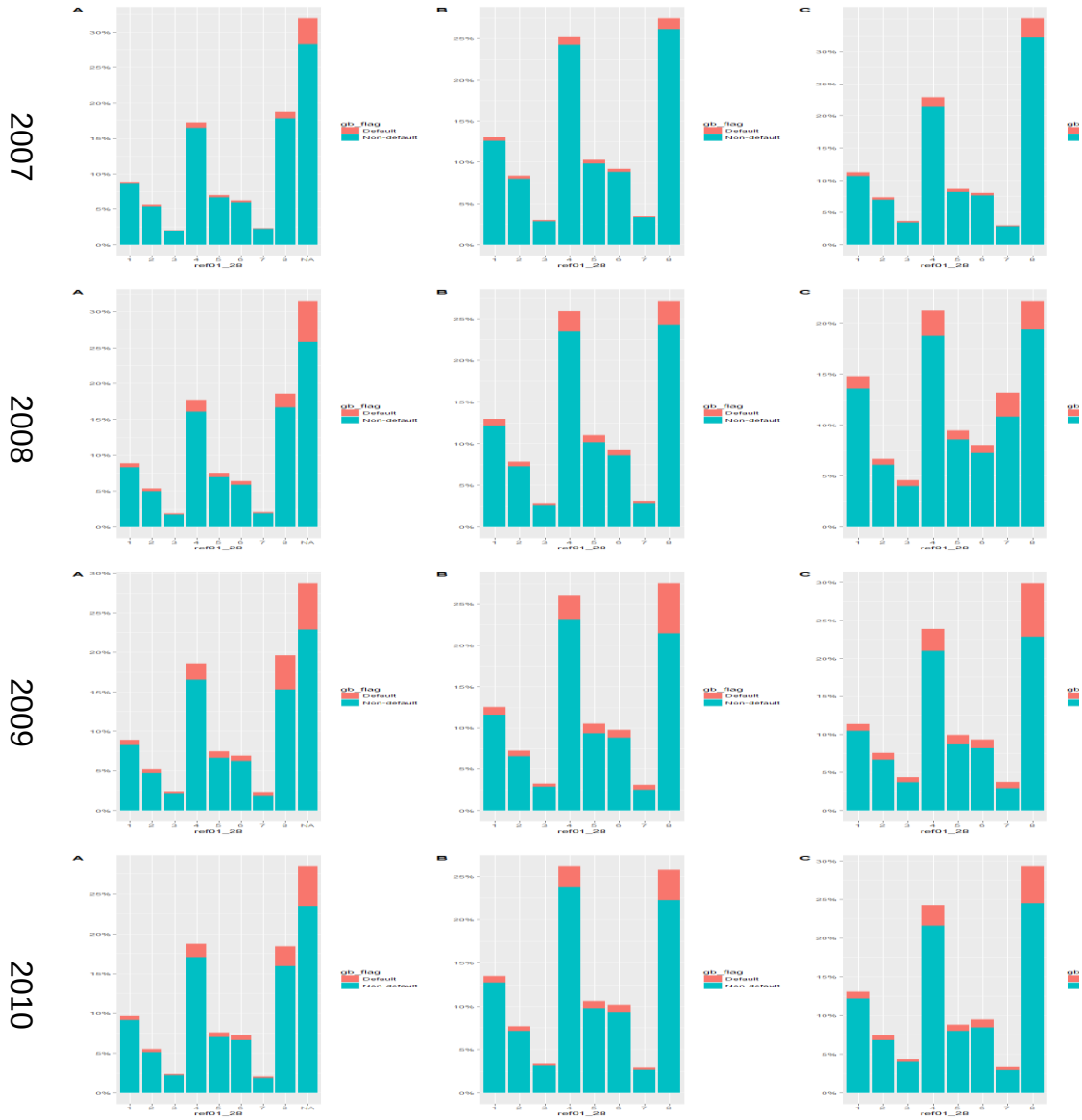
Overall, convergence for mean and standard deviation can be improved to make it more stable within a certain range without a clear trend. Visual inspection of imputation convergence leads to the undesirable choice of 20 burn-in iterations, although researchers suggested that simulation work using moderate amounts of

missing data yields satisfactory performance with just 5 or 10 iterations (Brand, 1999; Stef Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Similar density distributions are observed owing to using PMM. Besides, total fixed assets as a percentage of total assets has the highest FMI, averaging at 0.86, although it is not the one with the substantial missing values. FMI is not exactly same as missing values rate, and more importantly, not only variables with large amounts of missing data, but also that are weakly correlated with other variables in the imputation model would tend to have high FMI, thus leading trouble.

In the following, two categorical variables last derogatory item and 1992 SIC code are discussed because the former has the massive amount of missing values, while the latter is not significant during the financial crisis.

Figure 5-2 presents a series of bar charts to compare observed and imputed values of last derogatory item over years. As can be seen from the change of the Y-axis from panel A to panel B, this variable has substantial missing values which are classified as non-default. Once removing the missing values as shown in panel B, the fifth category occupies the enormous amount, especially at the peak of the financial crisis over 50% to default within this category. Panel C presents the distribution after imputation. Given a large amount of missing data, the shape of bar charts is obviously different from that of Pancel B. Specifically, the percentage of good versus bad in each bar changes. One of the distinct differences is that over 50% observations are non-default within the fifth category in 2009. This implies that it may produce biased estimates if solely focusing on the observed data and ignoring the missing data (listwise deletion).

Figure 5-3 lists bar charts from comparing observed and imputed values of variable 1992 SCI code throughout whole periods. Given a small number of missing values, both the shape of the observed and imputed distribution and the ratio of good against bad is similar among the panels. Because of the similar shape and ratio, the credit crisis had no major impact on the industrial classification.



Notes:

Imputed bar char is plotted by using pooled data over 50 imputations

The proportion of default (bad) versus non-default (good) is shown in each bar by colour.

A: bar chart of observed values keeping Missing values

B: bar chart of observed values removing Missing values

C: bar chart of imputed values

Figure 5-3 Bar chart of observed and imputed values of 1992 SIC code

5.1.2 Imputations of empirical variables

The whole sample is subdivided into start-ups and non-start-ups, and the variables used in imputation process has been determined based on (M. Ma, 2016). Likewise, continuous variables (nominal variables), binary variables and

categorical variables are imputed by PMM, logistic regression, and multinomial logistic regression respectively. Some researchers suggested the best method to impute limited-range variables is to impute on the raw scale with no restrictions to the range, and with no post-imputation rounding. Although this imputation method results in some implausible values, it appears to be the most consistent method with low bias and reliable coverage in repeated sampling of missingness, irrespective of the amount of skewness in the data (Von Hippel, 2013; Rodwell, Lee, Romaniuk, & Carlin, 2014). The imputation procedure produces an increasing to 100 imputed datasets with 50 maximum iterations for empirical variables with a fixed seed as well. Finally, Logistic regression is also used to pool estimates by Rubin's rules.

5.1.2.1 Pooling Results

The pooled estimates of selected variables of start-ups and non-start-ups are presented in Table 5-3 and Table 5-4 respectively. As mention early, the column FMI is the proportion of the total variance that is owing to the missing data (FMI = $(V_B + V_B /m)/V_T$).

As expected, the majority of variables are statistically significant. Yet, the contribution of 1992 SIC code is not during the whole observed periods for both segments, which is consistent to previous conclusions. Regions seems not to be a significant predictor for non-start-ups. However, there are unexpected findings. The high missing rate should come with high FMI. The pooled standard error should be higher than that from single imputed dataset because of between-imputation variance (V_B) and the extra variation due to the number of imputation (V_B/m). Variables: Proportion Of Current Directors To Previous Directors In The Last Year, Time since last derogatory data item (months), Time Since Last Annual Return, and Total Assets in start-ups, and Proportion Of Current Directors To Previous Directors In The Last Year, PP Worst (Company DBT - Industry DBT) In The Last 12 Months, Time since last derogatory data item (months), and Debt Gearing (%) in non-start-ups have extremely low standard error and coefficient estimates, given an extremely high FMI and missing rate over 50% (Table 3-2).

Table 5-3 Pooled results of Start-ups

Variables	2007			2008			2009			2010		
	est	se	fmi	est	se	fmi	est	se	fmi	est	se	fmi
(Intercept)	3.58	0.86	0.5	1.19	0.44	0.62	1.43	0.37	0.54	1.1	0.59	0.74
Legal form_1	-2.94	0.78	0.29	-2.23	0.65	0.62	-2.56	0.46	0.25	-2.91	0.85	0.55
Legal form_2	-3.64	0.52	0.18	-2.38	0.21	0.52	-2.41	0.15	0.3	-3.17	0.19	0.23
Legal form_3	-3.73	0.9	0.47	-2.16	0.63	0.53	-1.23	0.56	0.16	-2.23	0.59	0.2
Legal form_5	-2.59	0.69	0.38	-0.88	0.38	0.46	-0.72	0.31	0.44	-2.7	0.31	0.36
Legal form_6	9.68	93.58	0	12.02	98.47	0	11.82	94.9	0	11.38	90.42	0
Legal form_7	-2.73	0.63	0.35	-1.28	0.33	0.58	-1.28	0.22	0.35	-2.59	0.25	0.32
Legal form_8	11.07	320.55	0	13.35	393.8	0	13.7	461.22	0	12.45	534.15	0
Legal form_9										10.78	1455.4	0
Company is subsidiary_2	-0.25	0.24	0.71	1.11	0.2	0.7	1.69	0.12	0.32	1.18	0.17	0.45
Company is subsidiary_4	-0.35	0.74	0.9	0.18	0.58	0.59	2.63	0.53	0.19	1.25	0.54	0.45
1992 SIC code_2	2.08	112.18	0	-0.1	0.9	0.53	0.41	1.01	0.44	0.36	1.13	0.69
1992 SIC code_3	-0.55	1.08	0.77	-0.06	0.64	0.58	0.26	0.67	0.62	-0.79	0.8	0.7
1992 SIC code_4	-0.42	0.72	0.71	-0.59	0.39	0.63	0.11	0.32	0.53	0.08	0.57	0.8
1992 SIC code_5	-0.09	1.2	0.58	-0.62	0.81	0.65	0.45	0.66	0.56	0.24	0.78	0.72
1992 SIC code_6	-0.22	0.68	0.69	-0.37	0.36	0.62	0.24	0.3	0.52	0.08	0.55	0.81
1992 SIC code_7	-0.52	0.7	0.71	-0.2	0.36	0.63	0.04	0.3	0.53	0.04	0.55	0.8
1992 SIC code_8	-0.45	0.71	0.7	-0.38	0.39	0.65	0.12	0.31	0.52	0.07	0.57	0.8
1992 SIC code_9	-0.35	0.72	0.72	-0.29	0.37	0.61	-0.02	0.31	0.52	0.14	0.58	0.8
1992 SIC code_10	-0.36	0.79	0.74	-0.53	0.41	0.63	0.09	0.35	0.57	0.2	0.63	0.81
1992 SIC code_11	-0.37	0.69	0.71	-0.26	0.34	0.59	0.1	0.3	0.54	0.2	0.56	0.82
1992 SIC code_12	-0.43	0.7	0.72	-0.29	0.34	0.59	-0.01	0.3	0.55	0.11	0.55	0.81
1992 SIC code_13	-0.39	0.77	0.72	-0.18	0.44	0.64	0.14	0.36	0.53	0.26	0.61	0.8
1992 SIC code_14	-0.73	0.74	0.72	-0.52	0.4	0.65	0.09	0.32	0.54	0.26	0.56	0.79
1992 SIC code_15	-0.41	0.71	0.72	-0.22	0.36	0.62	0.1	0.3	0.53	0.17	0.56	0.81
Region_2	0.12	0.09	0.66	-0.12	0.07	0.53	0.22	0.06	0.29	0.24	0.07	0.5
Region_3	0.24	0.13	0.52	0.26	0.11	0.56	0.25	0.08	0.34	0.26	0.1	0.41
Region_4	0.09	0.21	0.52	0.32	0.19	0.52	0.18	0.13	0.28	0.13	0.17	0.5
Region_5	0.18	0.1	0.61	-0.48	0.08	0.62	0.12	0.06	0.27	0.12	0.08	0.41
Region_6	0.51	0.17	0.56	0.29	0.15	0.64	0.14	0.09	0.3	0.26	0.13	0.51
Region_7	0.25	0.1	0.48	0.18	0.1	0.6	0.34	0.07	0.33	0.1	0.08	0.4
Region_8	0.21	0.12	0.67	0.04	0.1	0.67	0.26	0.06	0.32	0.24	0.08	0.48
Region_9	-0.05	0.13	0.65	-0.18	0.1	0.54	0.24	0.08	0.33	0.06	0.09	0.45
Region_10	0.5	0.14	0.47	0.99	0.12	0.4	0.38	0.1	0.39	0.37	0.1	0.4
Region_11	0.39	0.21	0.6	-0.04	0.18	0.64	0.21	0.12	0.23	-0.2	0.15	0.48
Region_12	-1.25	0.81	0.52	-2.14	0.52	0.66	0.33	0.35	0.37	-1.41	0.34	0.64
Region_13	8.24	515.22	0	12.46	428.03	0	-0.59	1.41	0.12		475.06	0
Proportion of current directors to previous directors in the last year	0.27	0.1	0.92	0.14	0.09	0.94	-0.41	0.08	0.91	0.18	0.08	0.91
Oldest age of current directors /proprietors supplied (years)	0.01	0	0.73	0.01	0	0.72	0.02	0	0.46	0.01	0	0.53
Number of directors holding shares	0.11	0.06	0.72	0.93	0.05	0.66	1.14	0.04	0.48	0.78	0.04	0.53
Total value of judgements in the last 12 m	0	0	0.39	0	0	0.43	0	0	0.06	0	0	0.11
Number of previous searches (last 12m)	-0.03	0.02	0.74	0.01	0.02	0.69	0.05	0.01	0.35	0.08	0.01	0.4
Time since last derogatory data item (mo	0.42	0.03	0.88	0.39	0.03	0.93	0.37	0.03	0.91	0.42	0.03	0.89
Lateness of accounts	-0.15	0.01	0.71	-0.17	0	0.68	-0.18	0	0.5	-0.18	0	0.54
Time since last annual return	-0.15	0.01	0.77	-0.14	0.01	0.6	-0.15	0	0.45	-0.13	0.01	0.64
Total assets	0	0	0.91	0	0	0.9	0	0	0.84	0	0	0.79

Notes: est: estimate of coefficient; se: standard error; fmi: fraction of missing information
Orange: significant at 1% level, Yellow: significant at 5% level, Red: significant at 10% level

As indicated in literature review, one drawback of single imputation is to underestimate the standard error but multiple imputation doses not. It can verify by randomly selecting an imputed dataset (the 50th imputed dataset in 2009) and run a logistic regression to make comparison of the standard error for both start-ups and non-start-up. The regression results are shown in Appendix B, and it can conclude that standard errors from MICE are larger than that of the 50th imputed dataset from single imputation.

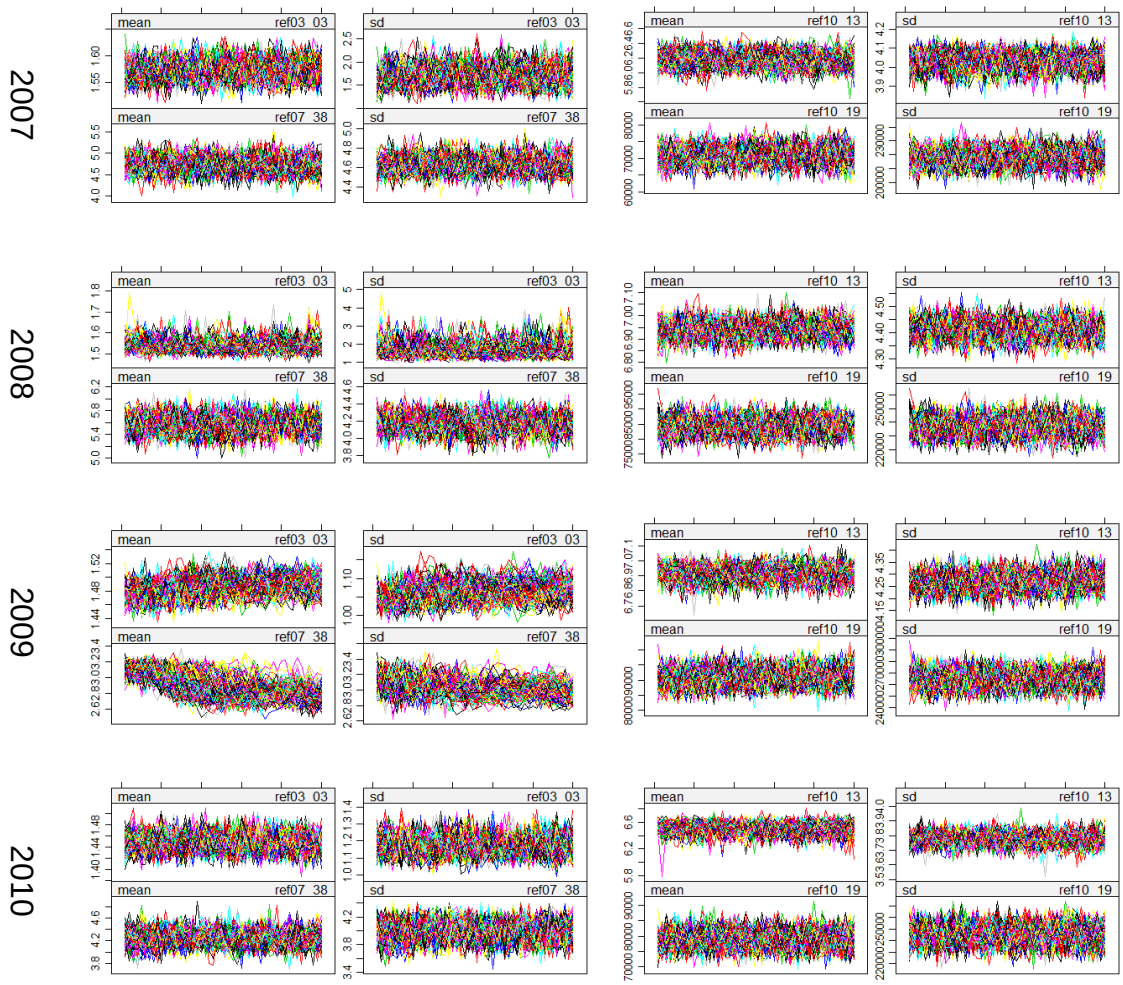
Table 5-4 Pooled results of Non-start-ups

Variables	2007			2008			2009			2010		
	est	se	fmi	est	se	fmi	est	se	fmi	est	se	fmi
(Intercept)	6.6	0.83	0.56	5.91	0.8	0.72	8.57	0.89	0.8	10.89	0.77	0.53
Legal form_1	-5.78	1.05	0.71	-4.09	0.79	0.53	-3.97	0.91	0.39	-9.62	1.15	0.66
Legal form_2	-4.35	0.69	0.55	-4.69	0.45	0.5	-6.47	0.52	0.61	-9.03	0.68	0.54
Legal form_3	-4.24	1.12	0.48	-3.16	1.64	0.63	-7.53	0.9	0.53	-10.16	1.15	0.6
Legal form_5	-4.13	0.9	0.62	-4.03	0.61	0.59	-6.64	0.64	0.64	-9.36	0.87	0.66
Legal form_6	17.54	143.87	0	17.78	220.19	0	16.93	83.87	0	15.57	123.08	0
Legal form_7	-3.84	0.8	0.58	-5.26	0.57	0.58	-6.35	0.56	0.58	-9.07	0.8	0.62
Legal form_8	17.48	188.74	0	21.14	259.56	0	8.71	1.59	0.45	21.68	151.97	0
Legal form_9	10.45	3956.2	0	11.3	6225.6	0	-8.58	3.78	0.22	5.47	3956.18	0
Parent company – derog details_2	-0.3	0.17	0.66	-0.24	0.16	0.75	-0.35	0.11	0.62	-0.03	0.11	0.53
Parent company – derog details_3	-1.55	0.63	0.43	-1.33	0.65	0.6	-0.23	0.33	0.29	-0.21	0.37	0.41
Parent company – derog details_4	-2.08	0.58	0.41	-0.61	0.7	0.35	-0.28	0.35	0.33	0.21	0.45	0.44
1992 SIC cide_2	-0.12	1	0.35	1.14	1.25	0.58	0.94	1.09	0.58	0.92	1.24	0.4
1992 SIC cide_3	-0.57	0.85	0.56	-1.4	1.29	0.87	-1.05	1.01	0.79	-1.07	0.8	0.69
1992 SIC cide_4	-0.52	0.43	0.54	0.17	0.53	0.72	-0.5	0.57	0.85	-0.7	0.36	0.52
1992 SIC cide_5	0.17	1.27	0.25	0.68	1.19	0.57	0.61	1.19	0.7	-1.16	1.02	0.7
1992 SIC cide_6	-0.26	0.4	0.49	0.2	0.52	0.73	-0.47	0.55	0.85	-0.49	0.36	0.53
1992 SIC cide_7	-0.38	0.41	0.52	0.03	0.51	0.72	-0.36	0.53	0.84	-0.42	0.37	0.57
1992 SIC cide_8	-0.33	0.42	0.49	-0.25	0.54	0.73	-0.79	0.56	0.85	-1.14	0.38	0.54
1992 SIC cide_9	-0.2	0.44	0.52	-0.11	0.54	0.72	-1.17	0.59	0.86	-0.61	0.38	0.54
1992 SIC cide_10	-0.27	0.52	0.56	0.04	0.56	0.69	-0.46	0.6	0.84	-0.54	0.45	0.64
1992 SIC cide_11	-0.27	0.39	0.49	0.06	0.5	0.71	-0.31	0.55	0.86	-0.56	0.35	0.54
1992 SIC cide_12	-0.2	0.39	0.48	0.22	0.49	0.7	-1.05	0.56	0.86	-0.77	0.34	0.51
1992 SIC cide_13	-0.53	0.55	0.56	-0.04	0.65	0.76	0.32	0.59	0.81	-0.3	0.46	0.6
1992 SIC cide_14	0.09	0.48	0.52	0.27	0.54	0.69	-0.46	0.61	0.84	-0.48	0.59	0.8
1992 SIC cide_15	-0.3	0.41	0.51	-0.32	0.5	0.7	-0.53	0.57	0.86	-0.54	0.42	0.66
Region_2	0.08	0.11	0.52	0.08	0.12	0.76	-0.51	0.08	0.65	0.03	0.1	0.66
Region_3	0.12	0.15	0.55	0.2	0.18	0.78	-0.07	0.12	0.65	-0.04	0.11	0.59
Region_4	0.19	0.28	0.62	0.51	0.26	0.75	0.05	0.18	0.6	-0.06	0.17	0.54
Region_5	0.07	0.13	0.51	0.29	0.15	0.75	-0.3	0.1	0.62	0.11	0.1	0.55
Region_6	0.36	0.18	0.43	0.63	0.16	0.61	0.01	0.15	0.66	0.15	0.15	0.63
Region_7	0.01	0.15	0.56	0.14	0.14	0.71	0.15	0.1	0.56	-0.26	0.1	0.56
Region_8	-0.12	0.14	0.58	-0.12	0.12	0.68	-0.19	0.1	0.59	0.06	0.09	0.49
Region_9	-0.06	0.18	0.63	-0.18	0.16	0.72	-0.51	0.1	0.55	0.05	0.12	0.58
Region_10	-0.09	0.18	0.6	0.83	0.17	0.66	0.27	0.12	0.56	0.44	0.11	0.48
Region_11	0.09	0.23	0.54	0.16	0.23	0.72	0.07	0.16	0.53	0.19	0.16	0.48
Region_12	-2.98	0.78	0.59	-1.56	0.6	0.58	-2.17	0.86	0.72	-8.7	0.73	0.58
Region_13	12.32	834.59	0	13.55	1352.7	0	14.61	478.87	0	8.57	1732744	0
No. Of 'current' directors	0.38	0.16	0.95	0.45	0.22	0.98	0.62	0.17	0.98	0.21	0.31	0.99
Proportion of current directors to Previous directors in the last year	-0.41	0.26	0.98	-0.65	0.4	0.99	-0.73	0.33	0.99	-0.08	0.62	1
Pp worst (company DBT - industry DBT) In the last 12 months	-0.01	0	0.8	0	0	0.82	-0.01	0	0.82	-0.01	0	0.82
Total value of judgements in the last 12 months	0	0	0.6	0	0	0.36	0	0	0.41	0	0	0.33
Number of previous searches (last 12m)	-0.01	0.01	0.53	-0.03	0.01	0.76	0	0.01	0.6	0	0.01	0.59
Time since last derogatory data item (months)	0.03	0	0.93	0.07	0.01	0.92	0.07	0.01	0.9	0.08	0.02	0.98
Lateness of accounts	-0.02	0	0.68	-0.03	0	0.71	-0.03	0	0.64	-0.04	0	0.73
Time since last annual return	-0.03	0.01	0.68	-0.05	0.01	0.77	-0.06	0	0.75	-0.05	0.01	0.85
Total fixed assets as a percentage of total assets	0	0	0.47	0.01	0	0.65	0.01	0	0.56	0.01	0	0.56
Debt gearing (%)	0	0	0.9	0	0	0.88	0	0	0.93	0	0	0.87
Percentage change in shareholders' funds	0	0	0.58	0	0	0.62	0	0	0.51	0	0	0.54
Percentage change in total assets	0	0	0.7	0	0	0.67	0	0	0.63	0	0	0.6

Notes: est: estimate of coefficient; se: standard error; fmi: fraction of missing information
Orange: significant at 1% level, Yellow: significant at 5% level, Red: significant at 10% level

5.1.2.2 Checking the Imputation Model

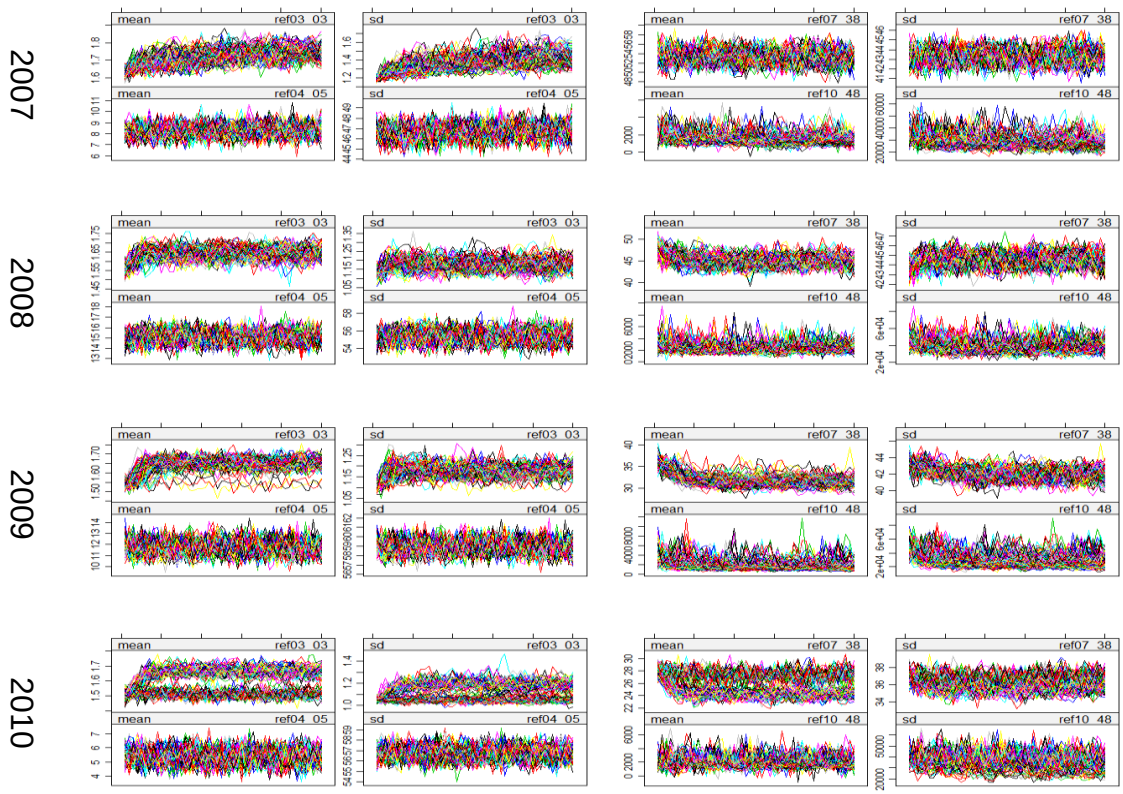
Figure 5-4 and Figure 5-5 provides convergence plots (the left one is mean, and the right one is standard deviation) for both start-ups and non-start-ups SMEs. In this section, the convergence plots of variables with over 50% missing rate are presented and discussed.



Notes: ref03_03: Proportion of current directors to previous directors in the last year; ref07_38: Time since last derogatory data item (months); ref10_13: Time since last annual return; ref10_19: Total assets

Figure 5-4 Convergence plots of variables over 50% missing rate for start-ups

Convergence results seem to be better as an increase of numbers of iteration and numbers of imputation. In 2009, the proportion of current directors to previous directors in the last year and time since last derogatory data item (months) in Figure 5-4 see an initial trend going upwards and downwards, respectively, and the trend remains till the end of the iteration.



Notes: ref03_03: Proportion of current directors to previous directors in the last year; ref04_05: Pp worst (company DBT - industry DBT) in the last 12 months; ref07_38: Time since last derogatory data item (months); ref10_48: Debt gearing (%)

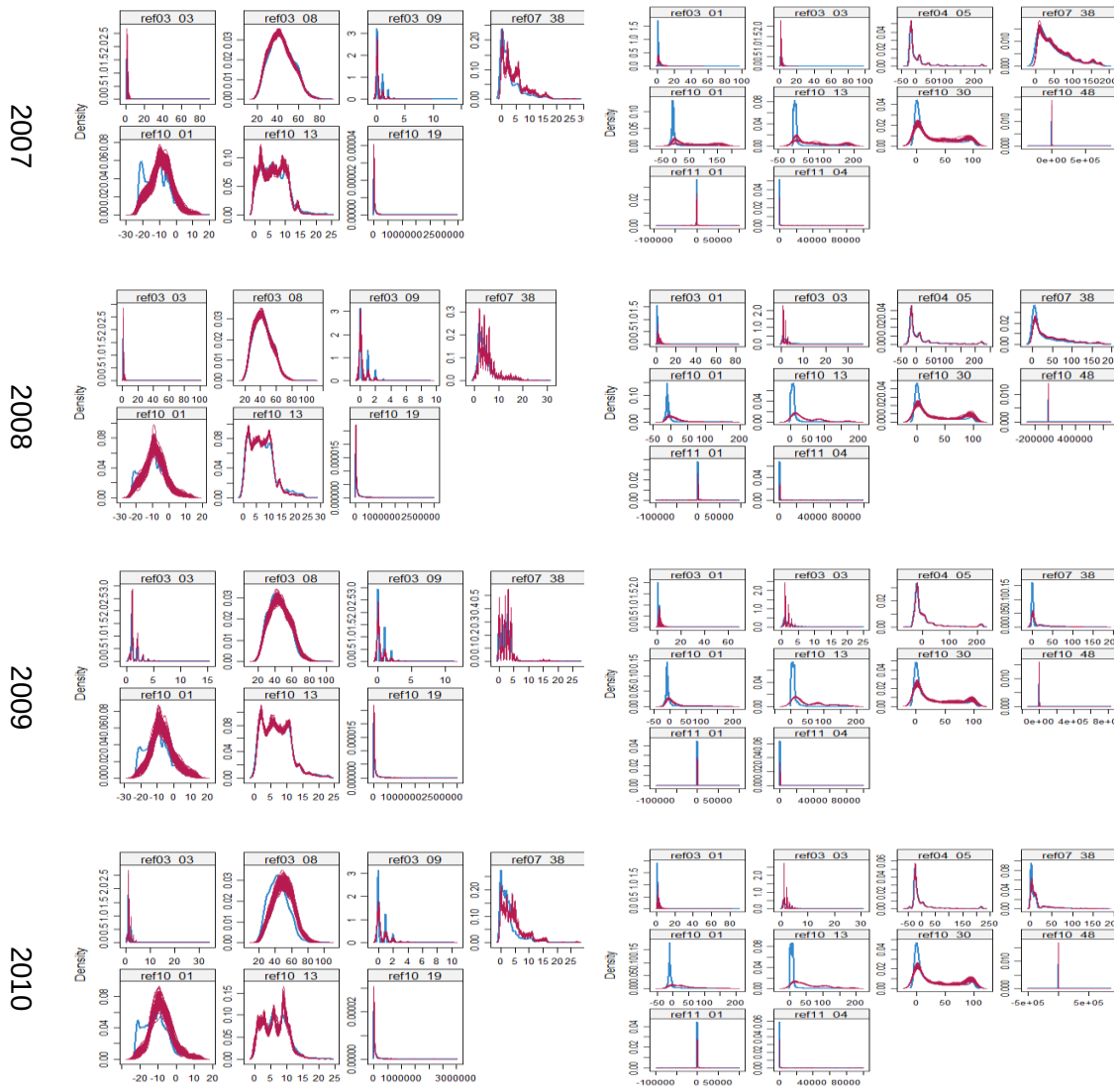
Figure 5-5 Convergence plots of variables over 50% missing rate for non-start-ups

In terms of non-start-ups variables (Figure 5-5), an initial trend can be found on variables proportion of current directors to previous directors in the last year and time since last derogatory data item (months), but the trend eventually remains stable. One of the worst convergences is proportion of current directors to previous directors in the last year in 2010. The plot shows a binary path. One path keeps stable since imputation begins, and another one remains static after an initial rising trend. Both paths do not converge until the end.

These problematic variables have an extremely high FMI, and it would be appropriate to increase the numbers of iteration or to require more correlated variables into the imputation process to obtain a more stable imputation results.

Figure 5-6 provides a series of density plots of missing variables from 2007 to 2010 for both start-ups and non-start-ups SMEs. As shown in the missing rate table (Table 3-2), variables have missing values ranging from less than 1% to up to 96.7%. On the left panel, imputed Lateness of accounts shows a “bell” shape, while the observed values are not. Due to less than 1% missing number, extreme values affect the shape of the plots dramatically. A similar situation can be found on the variable Number of directors holding shares. Another important variable needed to be checked is Time since last derogatory data item (months) because of its large missing rate. Intuitively, once this variable is greater than ten on the x-axis, observed and imputed values roughly overlap. Imputed values are approximately greater than the observed in the range of zero to ten. Imputed values are higher since short time since last derogatory data item is unwilling to report. Besides, impute values of other variables seem to have a good fit of the observed values.

On the right panel in Figure 5-6, imputation results seem to be much more problematic than that of start-ups. Missing rates of these variables: Proportion of current directors to previous directors in the last year, PP worst (company DBT - industry DBT) in the last 12 months, time since the last derogatory data item (months), and Debt gearing (%), are over 50%, and the others are less than 9%. There is a good fit between observed and imputed values of PP worst (company DBT - industry DBT) in the last 12 months and Debt gearing (%). The gap between observed and imputed values of Proportion of current directors to previous directors in the last year is found on the lower x-axis. As the proportion increases, the gap gradually disappears. It is different from start-ups of time since the last derogatory data item (months). Imputed values are smaller than observed values at a lower range of x-axis due to the beginning of the financial crisis.



Notes: The “blue” curve is generated by the observed value, and the “red” curve by imputed values from various imputed dataset. ref03_01: No. Of ‘current’ directors; ref03_03: Proportion of current directors to previous directors in the last year; ref03_08: Oldest age of current directors/proprietors supplied (years); ref03_09: Number of directors holding shares; ref04_05: Pp worst (company DBT - industry DBT) in the last 12 months; ref05_04: Total value of judgements in the last 12 months; ref06_03: Number of previous searches (last 12m); ref07_38: Time since last derogatory data item (months); ref10_01: Lateness of accounts; ref10_13: Time since last annual return; ref10_19: Total assets; ref10_30: Total fixed assets as a percentage of total assets; ref10_48: Debt gearing (%); ref11_01: Percentage change in shareholders’ funds; ref11_04: Percentage change in total assets.

Figure 5-6 Density plot of continuous variables

5.1.3 Variable Confirmation

In order to ensure a stable and consistent analysis results, it is necessary to use the same set of predictors over the years. It is reasonable to remove the effect in a specific year (significant only once). Apparently, this idea may identify different important variables with different thresholds, but there is no clear guide on how to choose a proper threshold. Either too much or too fewer variables would be selected as a result. Finally, considering both methods mentioned in Methodology chapter and the discussion in preceding sections, those variables significant three times or more are chosen as predictors, and then two sets of predictors are confirmed to further modelling the probability of default. The following table (Table 5-5) lists the variables used as independent variables to train the classifier from 'good' to 'bad' SMEs.

Table 5-5 Independent variables used for prediction

	Variables	Definitions
Start-ups	ref01_01	Legal form
	ref01_33	Region
	ref03_03	Proportion of current directors to previous directors in the last year
	ref03_08	Oldest age of current directors/proprietors supplied (years)
	ref03_09	Number of directors holding shares
	ref05_04	Total value of judgements in the last 12 months
	ref07_38	Time since last derogatory data item (months)
	ref10_01	Lateness of accounts
	ref10_13	Time since last annual return
	ref10_19	Total assets
Non-Start-ups	ref01_01	Legal form
	ref01_33	Region
	ref03_01	No. Of 'current' directors
	ref04_05	Pp worst (company DBT - industry DBT) in the last 12 months
	ref05_04	Total value of judgements in the last 12 months
	ref07_38	Time since last derogatory data item (months)
	ref10_01	Lateness of accounts
	ref10_13	Time since last annual return
	ref10_30	Total fixed assets as a percentage of total assets

5.1.4 Summary

In this section, the results of dealing with missing data using MICE are provided. The results of imputation are presented in three aspects: convergence, pooled result using logistic regression, and imputation check including various plots and static comparison between observed and imputed values. It is not fair to discard those missing observation. For example, time since last derogatory data item (months) with over 90% missing rate, but whichever method is applied, this variable always has a strong predictive power. There is numerous relationships among variables so that predictive power from missing data cannot be neglected.

The difference between method 1 (after-imputed) and methods 2 (empirical) are the following: method 1 makes use of information value (IV) to discard a small number of variables, and the rest of variables go into the imputation model. The imputation model includes a large number of undetermined variables, thus computation cost increases, but the influence of auxiliary variables are included. On the other hand, variables in method two have been determined previously. Second, regarding the gap between start-ups and non-start-ups, method 1 introduces a dummy variable while method two subdivides the whole sample. Hence, there are four imputation models for method 1 and eight models for method 2 given a four-year period. Third, the number of imputation and maximum iteration are different. In summary, the selection of predictive variables is of enormous difference.

According to the following aspects, a summary is provided.

1. Convergence: the judgment of convergence is an open question. Not all variables converge for both methods, yet method two is preferred since its convergence plots are more stable for variables with large missing rate (e.g., Time since last derogatory data item (months)). An increasing number of imputation and a maximum number of iterations do help improve convergence.
2. Pooled results: since segments dummy variables are significant in 2010 only (Table 5-1), segment effect may be ignored for method 1. On the other

hand, there are unexplained pooled results of method 2 for both segments. Coefficient and standard error are extremely small. Besides, regional effects after imputation are not significant for both methods, which is different from (M. Ma, 2016) selection. The only time since last derogatory data item (months) and time since last annual return are significant over every year and methods. Besides total fixed assets as a percentage of total assets is significant as well except in the start-ups' models of method 2. Both methods of FMI stay at a high level.

3. Imputation check: both methods use PMM to impute missing data, then the density of the majority of variables for observed and imputed values is similar. Using PMM may result in extreme values, so that affect the shape of density plots, especially for those variables with a small number of missing values. Generally, imputation diagnostic can be performed by making the comparison between the complete dataset and imputed dataset. Given the high volume of missing data in this research, it is not possible to perform a comparison to verify the accuracy of imputation since the observed data may be biased.

5.2 Cross-section Models

This section provides the results and findings of default prediction modelling after using imputed dataset and WoE data. Results of logistic regression using imputed dataset and WoE data will be first presented to explore the relation between dependent and independent variables from 2007 to 2010 since it is the benchmark model. After that, in order to prevent overfitting, results of shrinkage regression using WoE data is shown. Finally, the results of GAM is displayed to examine the non-linear behaviour of SMEs' performance further.

5.2.1 Logistic Regression with Weight on the Stacked Dataset

Generally, multiple imputed datasets are combined by Rubin's rules, and its estimated coefficient and standard errors have been shown in the last chapter. Yet

Rubin's rules may overestimate the standard error, and therefore a stacked data with weighted logistic regression could be a solution.

This section provides the results of stacked logistic regression with weights. There are dummy variables because of legal form and region. Coefficients and its standard error are shown for two segments Table 5-6 and Table 5-7. A general look at the results demonstrates that almost all variables for both segments are statistically significant at 99%. Table 5-6 presents the result for start-ups. Proportion of current directors to previous directors in the last year is one of the characteristics to describe board size in a SME. In the extremely changing economy situation under consideration this variable move from positive to negative sign, but it is still important. Although a larger board facilitates manager supervision and brings more human capital to advise managers, boards with too many members lead to problems of coordination, control, and flexibility in decision-making (De Andres and Vallelado, 2008). Therefore, a larger size of board may not able to lead a firm out of dilemma during the financial crisis in 2009. This finding is the same as number of appointments in the last 12 months as a percentage of the current board in the last section. Coefficient of Total value of judgement in the last 12 months in 2009 is not significant given a shock on the macroeconomic environment and its significance changes with time. Total assets is significant at 99% level yet it has an extremely low coefficient estimate and standard errors.

Table 5-7 provides the result of non-start-ups. Excluding Legal Form 6, Legal Form 8, and Legal Form 9, other categories are significant over the years. Legal Form 8, and Legal Form 9 are only significant in 2009. Majority of Region is not significant. No. of 'current' directors describe board structure in a firm as well. Its coefficient in 2009 suddenly increases from 0.13 to 0.36. This means an increase the number of directors is beneficial to lead the firm get through the crisis. The coefficient of Total value of judgement in the last 12 months sharply decreases since the outbreak of the credit crisis.

Table 5-6 Coefficients of logistic regression using a stacked dataset of start-ups with weights

	2007	2008	2009	2010
Legal Form 1	-3.000*** (0.800)	-2.000*** (0.490)	-2.400*** (0.480)	-2.900*** (0.690)
Legal Form 2	-3.700*** (0.570)	-2.200*** (0.170)	-2.200*** (0.150)	-3.000*** (0.200)
Legal Form 3	-3.700*** (0.790)	-1.900*** (0.540)	-1.200* (0.610)	-2.200*** (0.630)
Legal Form 5	-2.700*** (0.660)	-0.810** (0.340)	-0.710** (0.280)	-2.600*** (0.300)
Legal Form 6	10.000 (160.000)	12.000 (100.000)	11.000 (99.000)	11.000 (94.000)
Legal Form 7	-2.800*** (0.610)	-1.200*** (0.260)	-1.200*** (0.210)	-2.500*** (0.250)
Legal Form 8	12.000 (529.000)	13.000 (400.000)	14.000 (479.000)	13.000 (544.000)
Legal Form 9				11.000 (1,309.000)
Region South East	0.120* (0.065)	-0.140** (0.057)	0.210*** (0.056)	0.230*** (0.062)
Region South West	0.260** (0.110)	0.240*** (0.087)	0.240*** (0.081)	0.240*** (0.088)
Region North East	0.120 (0.180)	0.310** (0.160)	0.190 (0.130)	0.120 (0.140)
Region North West	0.200** (0.077)	-0.510*** (0.059)	0.076 (0.062)	0.110 (0.070)
Region East Midlands	0.520*** (0.130)	0.260** (0.110)	0.140 (0.093)	0.260** (0.110)
Region West Midlands	0.260*** (0.086)	0.150** (0.076)	0.290*** (0.068)	0.089 (0.077)
Region East England	0.230*** (0.082)	0.021 (0.069)	0.250*** (0.064)	0.230*** (0.070)
Region Yorkshire	-0.016 (0.088)	-0.220*** (0.080)	0.220*** (0.076)	0.060 (0.082)
Region Scotland	0.510*** (0.120)	0.950*** (0.120)	0.380*** (0.091)	0.320*** (0.096)
Region Wales	0.410** (0.160)	-0.063 (0.130)	0.170 (0.120)	-0.220* (0.130)
Region North Ireland	-1.200* (0.680)	-2.000*** (0.370)	0.420 (0.330)	-1.400*** (0.240)
Region Others	-0.110 (10.000)	12.000 (434.000)	-0.680 (1.500)	4.600 (10.000)
Proportion of current directors to previous directors in the last year	0.260*** (0.034)	**0.150*** (0.026)**	**-0.350*** (0.030)**	0.190*** (0.028)
Oldest age of current directors/proprietors supplied (years)	0.009*** (0.002)	0.010*** (0.002)	0.017*** (0.002)	0.012*** (0.002)
Number of directors holding shares	0.130*** (0.041)	0.880*** (0.035)	1.000*** (0.031)	0.730*** (0.035)
Total value of judgement in the last 12 months	-0.0001*** (0.00002)	-0.00003* (0.00002)	-0.00001 (0.00001)	-0.00004** (0.00002)
Time since last derogatory data item (months)	0.410*** (0.010)	0.390*** (0.009)	0.350*** (0.012)	0.420*** (0.010)
Lateness of accounts	-0.150*** (0.004)	-0.160*** (0.003)	-0.170*** (0.003)	-0.180*** (0.004)
Time since last annual return	-0.150*** (0.005)	-0.140*** (0.004)	-0.150*** (0.004)	-0.130*** (0.005)
Total Assets	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
Constant	3.300*** (0.580)	0.830*** (0.200)	1.300*** (0.170)	1.200*** (0.220)
Observations	3,311,700	3,312,610	3,034,710	2,830,800
Log Likelihood	0.000	0.000	0.000	0.000
Akaike Inf. Crit.	56.000	56.000	56.000	58.000
Note:		*p**p***p<0.01		

Table 5-7 Coefficients of logistic regression using a stacked dataset of non-start-ups with weights

	2007	2008	2009	2010
Legal Form 1	-5.000*** (0.670)	-4.000*** (0.640)	-4.900*** (0.820)	-9.400*** (0.780)
Legal Form 2	-3.800*** (0.540)	-4.600*** (0.380)	-6.700*** (0.400)	-8.800*** (0.540)
Legal Form 3	-3.400*** (0.990)	-2.600** (1.300)	-7.600*** (0.740)	-9.500*** (0.860)
Legal Form 5	-3.300*** (0.650)	-4.100*** (0.460)	-6.700*** (0.460)	-9.000*** (0.600)
Legal Form 6	17.000 (147.000)	18.000 (146.000)	18.000 (88.000)	16.000 (129.000)
Legal Form 7	-3.100*** (0.610)	-5.000*** (0.440)	-6.500*** (0.440)	-8.700*** (0.580)
Legal Form 8	17.000 (197.000)	21.000 (179.000)	9.400*** (1.500)	22.000 (159.000)
Legal Form 9	11.000 (3,779.000)	12.000 (3,766.000)	-8.500*** (2.100)	6.400 (3,577.000)
Region South East	0.084 (0.095)	0.067 (0.069)	-0.620*** (0.057)	0.056 (0.068)
Region South West	0.096 (0.120)	0.140 (0.098)	-0.008 (0.085)	-0.002 (0.084)
Region North East	0.110 (0.200)	0.320** (0.150)	-0.099 (0.130)	-0.022 (0.140)
Region North West	0.051 (0.110)	0.260*** (0.086)	-0.290*** (0.073)	0.100 (0.077)
Region East Midlands	0.340** (0.160)	0.530*** (0.120)	-0.036 (0.100)	0.150 (0.100)
Region West Midlands	-0.005 (0.120)	0.140 (0.092)	0.048 (0.078)	-0.260*** (0.077)
Region East England	-0.085 (0.110)	-0.093 (0.083)	-0.093 (0.074)	0.097 (0.076)
Region Yorkshire	-0.018 (0.130)	-0.170* (0.098)	-0.470*** (0.082)	0.078 (0.094)
Region Scotland	-0.046 (0.140)	0.760*** (0.120)	0.340*** (0.099)	0.390*** (0.096)
Region Wales	0.099 (0.190)	0.120 (0.150)	0.110 (0.130)	0.260* (0.140)
Region North Ireland	-2.400*** (0.580)	-1.600*** (0.470)	-2.100*** (0.570)	-8.400*** (0.550)
Region Others	1.700 (5.300)	5.000 (8.600)	18.000 (490.000)	-0.450 (5.900)
No. Of 'current' directors	0.140*** (0.035)	0.130*** (0.026)	0.360*** (0.026)	0.220*** (0.025)
Pp worst (company DBT - industry DBT) in the last 12 months	-0.007*** (0.0004)	-0.006*** (0.0003)	-0.006*** (0.0003)	-0.007*** (0.0003)
Total value of judgements in the last 12 months	-0.0001*** (0.00001)	-0.00002*** (0.00001)	-0.00003*** (0.00001)	-0.00003*** (0.00001)
Time since last derogatory data item (months)	0.025*** (0.001)	0.055*** (0.002)	0.061*** (0.002)	0.063*** (0.002)
Lateness of accounts	-0.021*** (0.003)	-0.029*** (0.003)	-0.034*** (0.002)	-0.039*** (0.003)
Time since last annual return	-0.028*** (0.004)	-0.052*** (0.003)	-0.062*** (0.003)	-0.051*** (0.003)
Total fixed assets as a percentage of total assets	0.004*** (0.001)	0.008*** (0.001)	0.010*** (0.001)	0.007*** (0.001)
Constant	5.700*** (0.560)	5.700*** (0.400)	7.700*** (0.410)	10.000*** (0.550)
Observations	2,870,280	3,235,330	3,437,350	3,470,740
Log Likelihood	0.000	0.000	0.000	0.000
Akaike Inf. Crit.	56.000	56.000	56.000	56.000
Note:			*p<0.1; **p<0.05; ***p<0.01	

5.2.2 Logistic Regression with WoE Data

In this section, results of logistic regression using original data with WoE transformation is presented. One of the powerful functions of WoE is able to handle missing data. Table 5-8 and Table 5-9 provide the coefficient estimation for start-ups and non-start-ups. For both segments, collinear variables have generally been removed, and no negative correlation exists.

Table 5-8 Coefficient estimates for logistic regression of start-ups data with woe transformation

	2007	2008	2009	2010
Legal Form	1.246*** (0.117)	1.599*** (0.095)	1.369*** (0.091)	1.404*** (0.115)
Region	0.628*** (0.106)	0.676*** (0.060)	0.495*** (0.091)	0.682*** (0.144)
Proportion of current directors to previous directors in the last year	0.625*** (0.092)	0.505*** (0.091)	0.517*** (0.110)	0.485*** (0.149)
Oldest age of current directors/proprietors supplied (years)	0.622*** (0.088)	0.705*** (0.036)	0.715*** (0.066)	0.735*** (0.077)
Number of directors holding shares	0.574*** (0.092)	1.020*** (0.068)	1.112*** (0.092)	1.251*** (0.147)
Total value of judgement in the last 12 months	0.412*** (0.104)	0.244 (0.151)	0.612*** (0.096)	0.497*** (0.101)
Time since last derogatory data item (months)	0.714*** (0.026)	0.646*** (0.024)	0.703*** (0.017)	0.762*** (0.020)
Lateness of accounts	1.381*** (0.043)	1.074*** (0.024)	0.825*** (0.020)	0.942*** (0.027)
Time since last annual return	1.234*** (0.043)	0.899*** (0.029)	0.759*** (0.033)	0.941*** (0.047)
Total Assets	0.735*** (0.069)	0.499*** (0.056)	0.604*** (0.044)	1.147*** (0.061)
Constant	2.139*** (0.022)	1.511*** (0.018)	1.243*** (0.018)	1.514*** (0.019)
Observations	33,117	33,126	30,347	28,308
Log Likelihood	-7,883.157	-10,658.430	-10,526.470	-9,632.197
Akaike Inf. Crit.	15,788.310	21,338.860	21,074.940	19,286.390

Note: *p<0.1; **p<0.05; ***p<0.01

For start-ups, Total value of judgement in the last 12 months significance changes when the economy changes. It loses significance at the beginning of credit crisis and recovers significance at the peak of credit crisis.

Non-start-ups in Table 5-9, one noticeable difference is found on variable Region. In the general economic cycle, regional differences are not obvious. Once the credit crisis begins, the importance of regional differences starts to emerge and continues to recover period. Significance of other variables remain stable, and its coefficients clearly change over the years.

Table 5-9 Coefficient estimates for logistic regression of non-start-ups data with woe transformation

	2007	2008	2009	2010
Legal Form	0.816*** (0.168)	0.988*** (0.125)	0.563*** (0.076)	1.426*** (0.122)
Region	0.506* (0.296)	0.560*** (0.133)	0.572*** (0.053)	0.802*** (0.128)
No. Of 'current' directors	0.562*** (0.097)	0.537*** (0.071)	0.499*** (0.040)	0.547*** (0.062)
Pp worst (company DBT - industry DBT) in the last 12 months	0.606*** (0.097)	0.356*** (0.065)	0.309*** (0.042)	0.560*** (0.058)
Total value of judgements in the last 12 months	0.484*** (0.061)	0.374*** (0.072)	0.327*** (0.080)	0.348*** (0.062)
Time since last derogatory data item (months)	0.491*** (0.046)	0.615*** (0.022)	0.631*** (0.015)	0.703*** (0.017)
Lateness of accounts	0.671*** (0.037)	0.597*** (0.025)	0.446*** (0.020)	0.493*** (0.026)
Time since last annual return	0.749*** (0.039)	0.658*** (0.024)	0.530*** (0.018)	0.485*** (0.027)
Total fixed assets as a percentage of total assets	0.819*** (0.098)	0.814*** (0.049)	0.712*** (0.029)	0.748*** (0.050)
Constant	2.938*** (0.031)	2.263*** (0.024)	1.655*** (0.021)	2.141*** (0.022)
Observations	28,702	32,353	34,373	34,707
Log Likelihood	-4,729.743	-7,190.717	-8,637.401	-7,813.772
Akaike Inf. Crit.	9,479.486	14,401.430	17,294.800	15,647.540
Note:		*p<0.1; **p<0.05; ***p<0.01		

5.2.3 Shrinkage Regression with WoE Data

A popular and successful approach in statistical modelling is to use regularization penalties in model fitting. By jointly minimizing the empirical error and penalty, one seeks a model that not only fits well and is also simple avoiding considerable variation which occurs in estimating complex models. Both ridge and lasso produce a more "regularized" model when compared with logistic regression. Ridge regression will not remove any variable but minimize the coefficient estimated, and lasso are more interpretable since there may be zero coefficient, and therefore, lasso regression has a function of variable selection.

To determine the lambda, 10-fold cross-validation is used, and as explained before, lambda.1se is selected to produce coefficient and predict accuracy. Higher the value of lambda, greater will be the shrinkage of the coefficients and this, in turn, makes the coefficients more robust to collinearity.

Table 5-10 Coefficient estimates for start-ups for lambda.1se

Variables	2007		2008		2009		2010	
	Lasso λ=0.007	Ridge λ=0.3584	Lasso λ=0.0152	Ridge λ=0.0833	Lasso λ=0.0129	Ridge λ=191.7774	Lasso λ=0.0089	Ridge λ=157.743
(Intercept)	2.0915	2.1488	1.4557	1.4549	1.2319	1.3415	1.4851	1.5671
Legal Form	0.4561	0.0884	0.5386	0.5364	0.5835	0.0006	0.8489	0.0007
Region	0.2265	0.1467	0.2734	0.4052	0.0853	0.0008	0.3025	0.0009
Proportion of current directors to previous directors in the last year	0.1865	0.1133	-	0.334	-	0.0006	0.0968	0.0006
Oldest age of current directors/proprietors supplied (years)	0.3835	0.1636	0.5824	0.5096	0.3852	0.0009	0.6177	0.0009
Number of directors holding shares	0.1669	0.1253	0.4914	0.4163	0.4874	0.0007	0.2917	0.0003
Total value of judgement in the last 12 months	0.1456	0.2585	-	0.3945	0.1089	0.0011	0.2747	0.0013
Time since last derogatory data item (months)	0.6785	0.325	0.5305	0.4881	0.6427	0.001	0.7293	0.0013
Lateness of accounts	1.0168	0.1733	0.7742	0.4769	0.635	0.0007	0.7654	0.0007
Time since last annual return	0.9231	0.1348	0.6795	0.4682	0.565	0.0007	0.7771	0.0008
Total Assets	0.393	0.1108	0.1502	0.3041	0.3856	0.0006	0.9556	0.0007

Table 5-11 Coefficient estimates for non-start-ups for lambda.1se

Variables	2007		2008		2009		2010	
	Lasso λ=0.0041	Ridge λ=38.3729	Lasso λ=0.0096	Ridge λ=107.3574	Lasso λ=0.0095	Ridge λ=181.1773	Lasso λ=0.0089	Ridge λ=144.7594
(Intercept)	2.8915	2.9624	2.2036	2.3001	1.6235	1.7353	2.1053	2.1793
Legal Form	0.3213	0.0008	0.3075	0.0005	0.1872	0.0004	0.4806	0.0004
Region	-	0.0013	-	0.0007	0.3241	0.0007	-	0.0007
No. Of 'current' directors	0.223	0.0005	0.1128	0.0004	0.2587	0.0003	0.1939	0.0003
PP worst (company DBT - industry DBT) in the last 12 months	0.3285	0.001	0.0394	0.0006	0.1084	0.0005	0.2535	0.0005
Total value of judgement in the last 12 months	0.4284	0.0026	0.1019	0.0012	-	0.0009	0.1541	0.0011
Time since last derogatory data item (months)	0.4782	0.0034	0.5763	0.0016	0.5969	0.001	0.7066	0.0012
Lateness of accounts	0.629	0.0017	0.5308	0.001	0.3946	0.0008	0.4123	0.0008
Time since last annual return	0.6788	0.0017	0.5694	0.0009	0.4903	0.0008	0.3749	0.0008
Total fixed assets as a percentage of total assets	0.4832	0.0013	0.551	0.0009	0.5788	0.0008	0.4246	0.0008

For start-ups (Table 5-10), it is suggested that Proportion of current directors to previous directors in the last year may be not a significant variable during the credit crisis, and this indicates structure of mangers has no great impact on the performance of SMEs during the credit crunch. For non-start-ups (Table 5-11), excluding year 2009, it is suggested that region should be removed from the prediction model, and in 2009, total value of judgement in the last 12 months should be discarded as well. Lambda of ridge regression for both segments

sharply rise during and after the credit crisis. The Ridge regression does not produce zero estimates even for large values of lambda.

5.2.4 Generalised Additive Model (GAM) with Imputed Dataset

The GAM models were established based on the response variable and its significant and influential predictors using the imputed dataset. The specific importance and effect of each predictor imposed on the response can be examined from the GAM results. The relative significance of each predictor can be quantified and compared by the significance (p-value) associated with each smoothed term in the GAM. The effect of each predictor on the response can be described by the effective degree of freedom (EDF) and the function plot of each smoothed term.

Contributions for predictor variables in the GAM and the effect of each predictor can be partitioned and examined by the smooth function plots. The plot presents the varying magnitude of the effect of each variable where the y-axis represents the contribution (effect) of each covariate to the fitting response, centred on zero. The numbers in the labels of the y-axis denote the effective degrees of freedom. The relative density of data points is shown by the rug plot on the x-axis. Rug plots are particularly useful in connection with additive models where the plotted smooth function is used to assess how much data contributed to the model fit at the different values of the independent variables. Estimated smooth functions (solid lines) with 95% confidence intervals (shaded area) are shown for each predictor. The positive slope of the smoothed line indicates a positive effect of the predictor imposed on the ‘good’ estimation and vice versa. The narrow confidence limits indicate high relevance and wide confidence limits indicated low relevance ranges of distribution (Solanki, Bhatpuria, & Chauhan, 2016).

If the smoothing leads to be a linear model then it consequently has one degree of freedom, and there is no choice about where it passes through zero. Therefore, the confidence interval must vanish at that point. If the confidence intervals had overlapped with zero for certain values of x (or throughout the entire range), this

would imply a non-significant effect at those x values (or of x in entirety) when the contribution for individual variable changes along the range of x-axis, the change in that covariate is associated with a change in the response.

5.2.4.1 Start-ups

Table 5-12 lists the effective degree of freedom and approximate significance associated with each smoothed term in the final GAM model for start-ups from 2007 to 2010. The result shows that all the 8 variables included in the GAM are statistically significant (p-value <0.05) and thus essential in the prediction model. Of the 8 independent variables included in the GAM, Time since last derogatory data item (months), Lateness of accounts, Time since last annual return, and Total assets present the highest influential behaviour during the whole observed periods (smallest values of p-value). Most of the predictor variables exhibit a significant non-linear effect for modelling (EDF >1), excluding Proportion of current directors to previous directors in the last year in 2007, 2009 and 2010, Total value of judgements in the last 12 months in 2007 and 2008 in start-ups (EDF ≈ 1). In the following, smooth function plots of start-ups (Figure 5-7 - Figure 5-14) are discussed in detail by variables.

- Proportion of current directors to previous directors in the last year

As shown in Figure 5-7, the non-linear trend only observes in 2008 as the EDF is approximately 5, and the linear trend is observed in the rest of the years. Over the observed period, a thin confident band is observed from around 1 to 3. It can be concluded from the rug plot that majority observations are smaller than 5. Confidence limit is especially narrow when around zero, and it expands sharply especially in 2008.

Table 5-12 Effective degrees of freedom and approximate significance of each GAM smoothed term of start-ups

Smoothed term	EDF				p-Value			
	2007	2008	2009	2010	2007	2008	2009	2010
Proportion of current directors to previous directors in the last year	1	5.06	1.02	1.05	8.4×10^{-3}	2.00×10^{-16}	2.00×10^{-16}	2.43×10^{-10}
Oldest age of current directors/proprietors supplied (years)	2.6	3.05	2.8	2.58	8.23×10^{-12}	5.79×10^{-11}	2.00×10^{-16}	2.00×10^{-16}
Number of directors holding shares	3.85	4.35	4.66	4.3	5.27×10^{-9}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Total value of judgements in the last 12 months	1.02	0.81	4.29	5	5.08×10^{-7}	3.32×10^{-2}	5.26×10^{-5}	3.78×10^{-5}
Time since last derogatory data item (months)	7.33	6.53	5.91	6.73	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Lateness of accounts	7.5	7.37	14.7	7.43	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Time since last annual return	6.33	6.13	7.36	6.43	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Total assets	15.2	5.53	7.34	5.38	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}

Notes: 2009 model uses a small ridge penalty added to the smoothing penalty so that the whole term

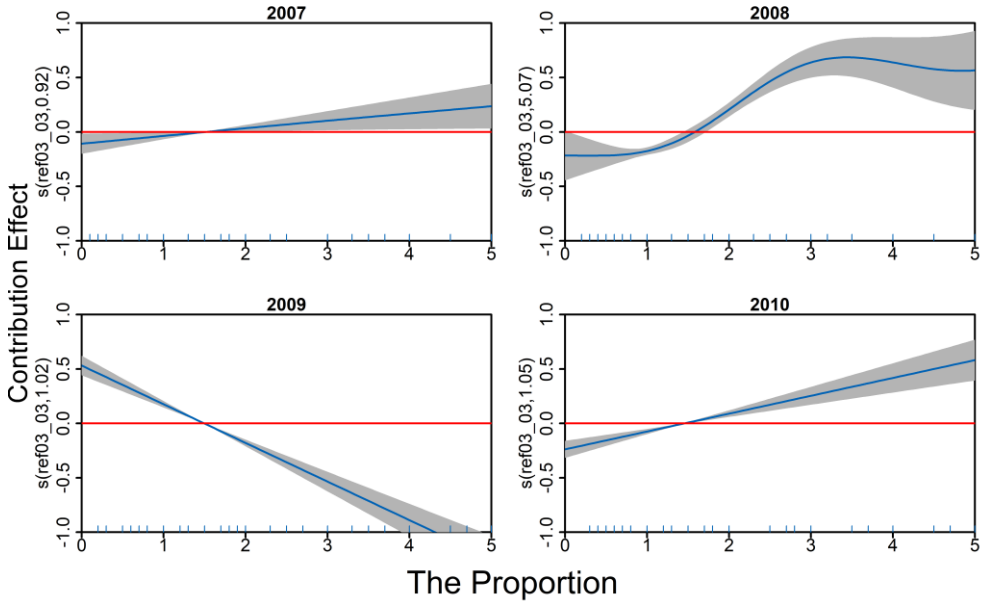


Figure 5-7 GAM - Proportion of current directors to previous directors in the last year

One major difference is found at the negative slope in 2009, while others show a positive slope. This means increasing the proportion would lead to a decrease of

the probability of being ‘good’ during the peak of the financial crisis; more new directors (larger size of the board) could not tide over the crisis. Changes in the economic environment have a great impact on this variable.

Negative coefficient shows that there is a point at which adding a new director reduces bank value. Board size increases with the firm’s development, and it reflects a trade-off between the firm-specific benefits and costs of monitoring (Boone, Casares Field, Karpoff, & Raheja, 2007). Boards with many directors are able to assign more people to supervise and advise on managers’ decisions. Having more supervisors and advisors either reduces managers’ discretionary power or at least makes it easier to detect managers’ opportunistic behaviour. Besides, it increases strategic capabilities to complement that of the CEO, up to a certain limit (De Andres and Vallelado, 2008). Aebi, Sabato, & Schmid (2012) found that during the credit crisis, board characteristics that are usually considered good corporate governance were mostly insignificantly or even negatively related to bank performance. With regard to board size, it is adversely related to the bank’s performance. Guest (2009) found the same relation for UK listed firms over 1981–2002 in which there was a financial crisis. Yet this finding could be argued by the difference between banks and SMEs.

- Oldest age of current directors/proprietors supplied (years)

The smooth function plot (Figure 5-8) can be divided into three parts for discussion. The first part is those directors younger than 40. They are young and can provide new ideas or strategies to help companies develop yet could be a lack of experience. All smooth functions present a positive trend, but the narrowest confidence band is observed in 2009. The second part is the directors’ age from 40 to 60. Similarly, all smooth functions go upwards but at a slower rate, and the confidence band becomes much narrower. The third part is directors older than 60. Negative influence with a border confidence band can be observed in 2008, while positive influence with a wide confidence band can be observed in 2009. There is a relatively flat curve in 2007 and 2010. From rug plots, observations, especially from directors older than 80, are insufficient, thus leading not significant with zero.

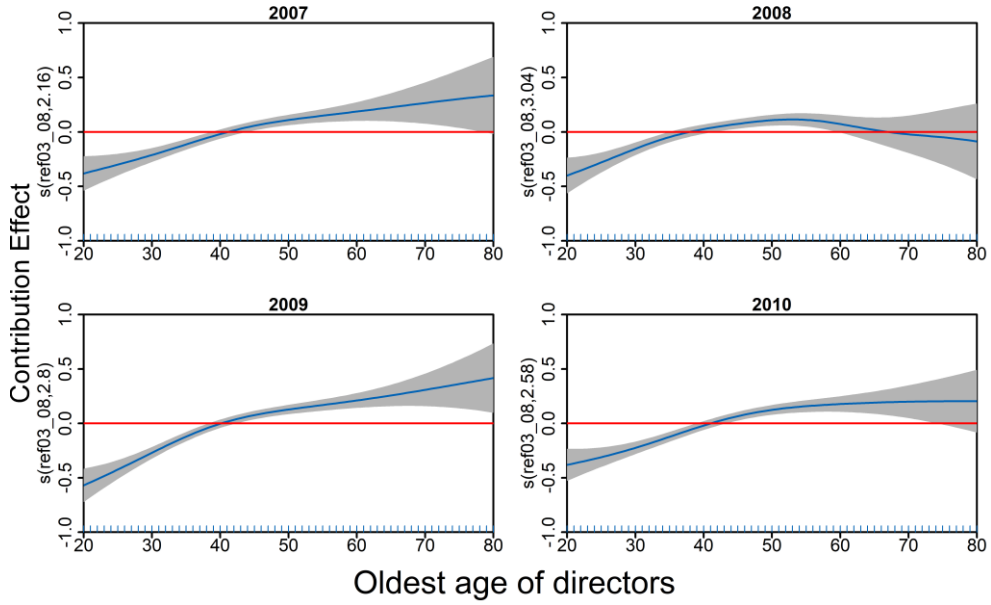


Figure 5-8 GAM - Oldest age of current directors/proprietors supplied (years)

Directors gain more knowledge and practical experience as they age. This could boost SMEs' performance and is beneficial to survive especially during the peak of the credit crisis. However, this beneficial influence may disappear gradually as the directors' age increase. In addition, changes in the economic environment have less impact on this variable due to a consistent pattern of the curve through the years.

- Number of directors holding shares

Shareholders and directors have two completely different roles in a company. The shareholders own the company by owning its shares and the directors manage it. A director does not need to be a shareholder and a shareholder has no right to be a director. It is not surprising that one has a dual role of director and shareholder in SMEs due to the number of employees.

The smooth function plot (Figure 5-9) can be divided into three parts for discussion. The first part is from zero to one. With an extremely narrow confidence band, there is an overall increasing trend. This indicates that one person with dual roles (director and shareholder) is good for the firm's development at any period. The second part is from one to four. Generally, the smooth curve is relatively flat with

limited fluctuations in 2009, while other curves display a slightly negative movement. The third part is above four. This part with the largest confidence limits mainly contains outliers, thus making uncertainty and less predictable. Changes in the economic environment have a mild impact on this variable.

Beltratti and Stulz (2012) concluded that banks with more shareholder-friendly boards performed significantly worse during the crisis. This may indicate that banks were pushed by their boards to maximize shareholder wealth before the crisis and took risks that were understood to create wealth but later turned out poorly during the credit crisis (Aebi, et al., 2012).

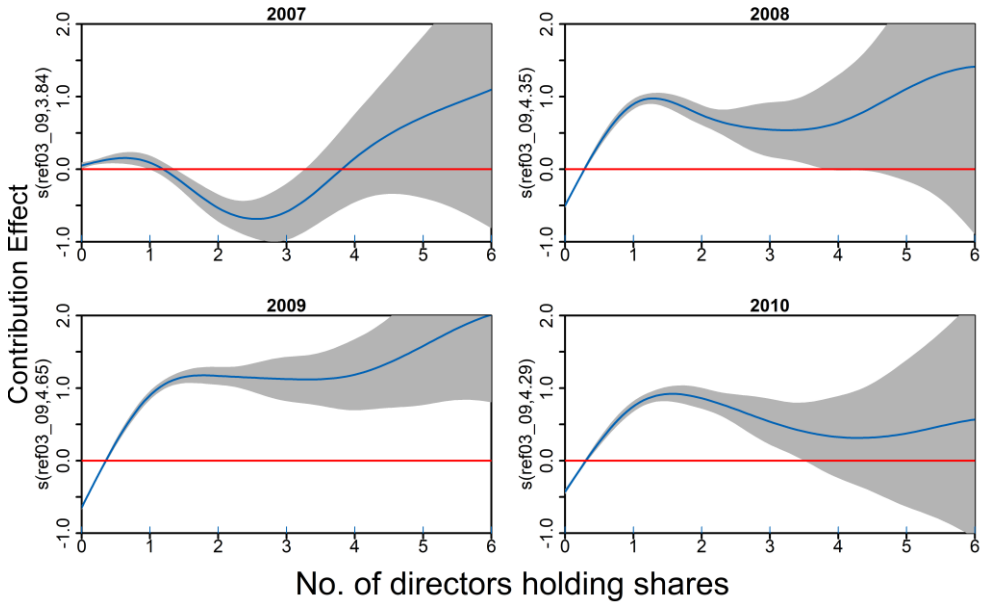


Figure 5-9 GAM - Number of directors holding shares

The smooth function plot (Figure 5-9) can be divided into three parts for discussion.

The first part is from zero to one. With an extremely narrow confidence band, there is an overall increasing trend. This indicates that one person with dual roles (director and shareholder) is beneficial for the firm's development at any period. The second part is from one to four. Generally, the smooth curve is relatively flat with limited fluctuations in 2009, while other curves display a slightly negative movement. The third part is above four. This part with the wide confidence limits

mainly contains outliers, thus making uncertainty and less predictable. Changes in the economic environment have a mild impact on this variable.

In summary, a small number of directors holding shares is good for SMEs development. Yet, with the increase of that number, this impact would not increase simultaneously. At the peak of the credit crisis, although more than two directors holding shares will not necessarily bring benefits, but an increasing number of that certainly will not bring harm.

- Total value of judgements in the last 12 months

If the obligator is not paying loans back, a judge would be made with regard to the unsettled loan. The judgement record would persist even if payment was made after the judgement (M. Ma, 2016). Therefore, it is expected that SMEs with better performance has a lower value of judgements. Since the measurement time is last 12 months, it is not surprising that the same pattern is observed in 2007 and 2008 because its measurement happens in the normal period. Hence, observations in 2009 correctly record the values since the outbreak of credit crisis, and observations in 2010 record the values at the peak of the credit crisis. It can thereafter conclude that switch of the economic environment has a substantial impact on this variable because of the shape of curves (Figure 5-10).

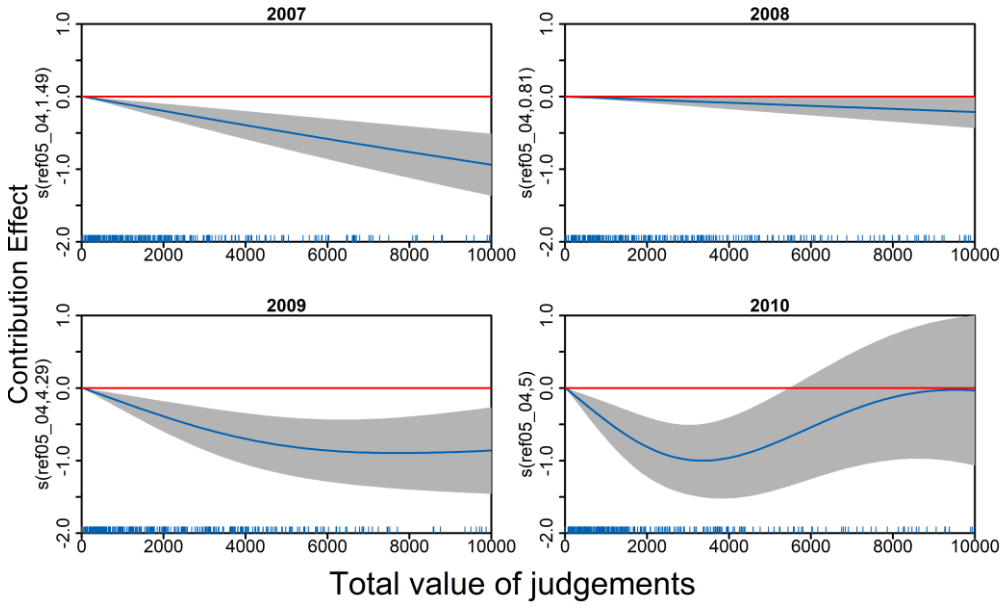


Figure 5-10 GAM - Total value of judgements in the last 12 months

The figure in 2009 (Figure 5-10) presents the initial trend from zero to fifty thousand. Non-linear relation is found in 2009 and 2010. The initial negative trend is similar, after that smooth functions in 2009, and 2010 with wide confidence limits become more and more volatile especially in 2010. A positive effect can be observed between 2009 and 2010. It can conclude from a rug plot that board confidence band may result from a lack of data in the variables.

In summary, a strongly adverse relation is identified in 2007 and 2008, but a relatively weak adverse relation is identified in 2009 and 2010.

- Time since last derogatory data item (months)

The smooth function plot (Figure 5-11) can be divided into three parts for discussion. The first part is from zero to six months. An initially flat curve is found in 2008, but an overall positive influence is presented from 2007 to 2010 in general. This impact is very significant because of a very narrow confidence band. The curve in 2007 is steeper than others. The second part is from six to twenty months. In general, the curve is still climbing, though, at a lower rate. Two peaks are observed in 2007, while other curves are relatively flat. The third part is above twenty months. This variable has no significant influence in 2010 due to large confidence limits, and its curve is relatively flat. Other curves still show a rising trend with reasonable confidence bands.

The derogatory data is especially important if the record is more recent. Start-ups with a recent derogatory record significantly jeopardize SMEs' performance, but this effect is relatively weak after the sixth month.

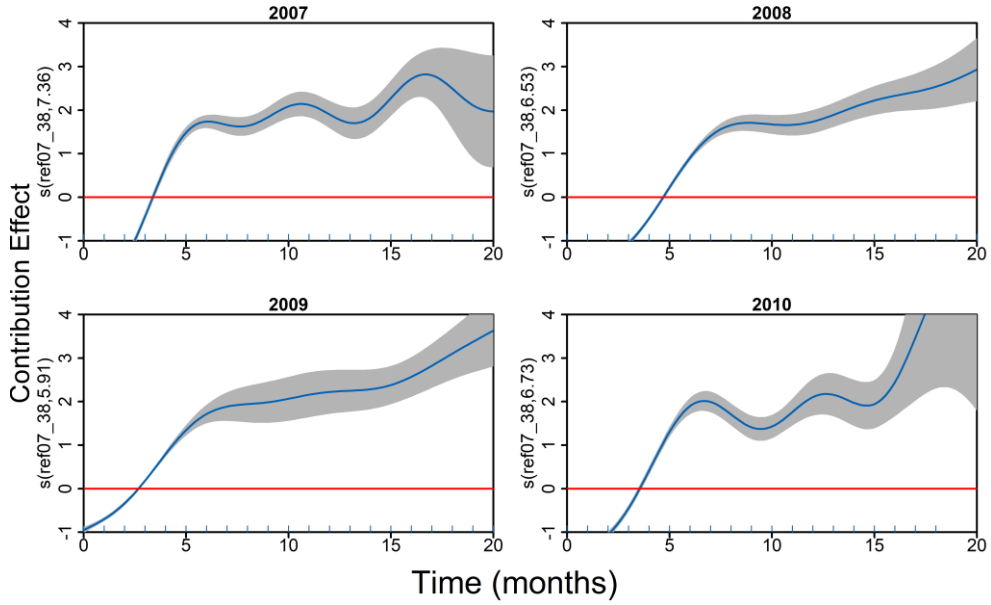


Figure 5-11 GAM - Time since last derogatory data item (months)

- Lateness of accounts

Generally, an inverse trend is observed from 2007 to 2010 (Figure 5-12). The smooth function can be divided into three parts to discuss. The first part is below -20. The curve shows a negative trend, but the curve in 2007 and 2009 are steeper. The second part is from -20 to 10. Negative influence with narrow confidence band is observed, although there are apparent fluctuations in 2007. SMEs performance decrease as the time since the last accounting update become longer. The third part is above 10. Except in 2010, an increasing trend with wide confidence bands is found from 2007 to 2010.

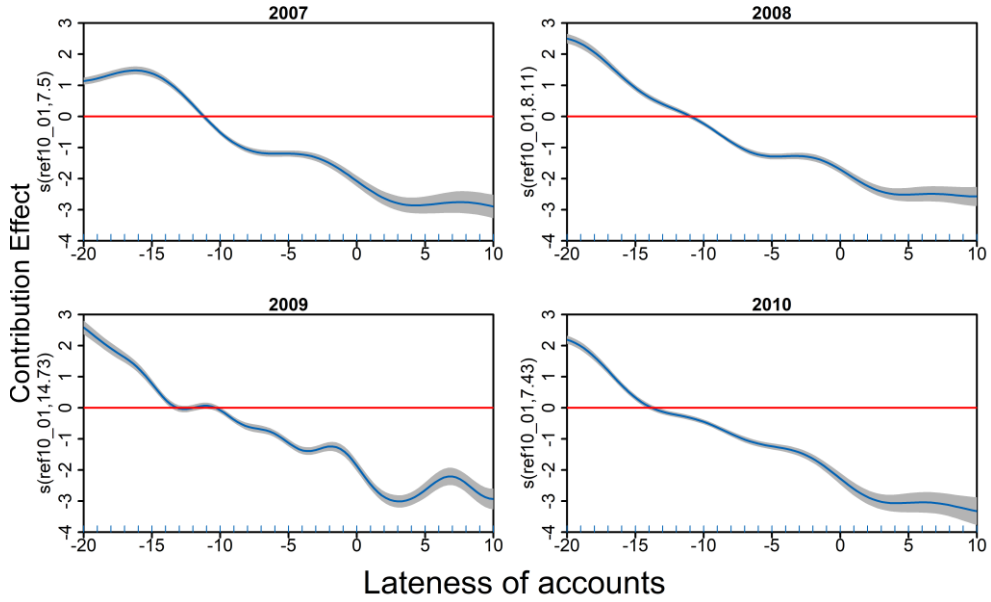


Figure 5-12 GAM - Lateness of accounts

In order to accurately capture the trend, a higher EDF is to be estimated in 2009. Considering the stable shape of the overall trend, it can conclude that the business cycle would not affect this variable significantly, and it might be regarded as generally linear and consistent.

- Time since last annual return

Likewise, a decreasing trend is observed from 2007 to 2010 (Figure 5-13) in general. The part above 20 needs to be further discussed. Above this part, the smooth functions with narrow confidence limits go downward significantly with a flat near 10. The smooth curve in 2007 increases with a wide confidence band, while other curves decrease at a decreasing rate.

Time since last annual return marks the duration since the last time the firm reported to Companies House. The shorter the time since the last reporting, the more transparent the SME's information. This is a very strong conclusion, which can help banks separate 'good' SMEs from 'bad' according to the punctuality of their annual return even if SMEs do not provide detailed 'soft' information.

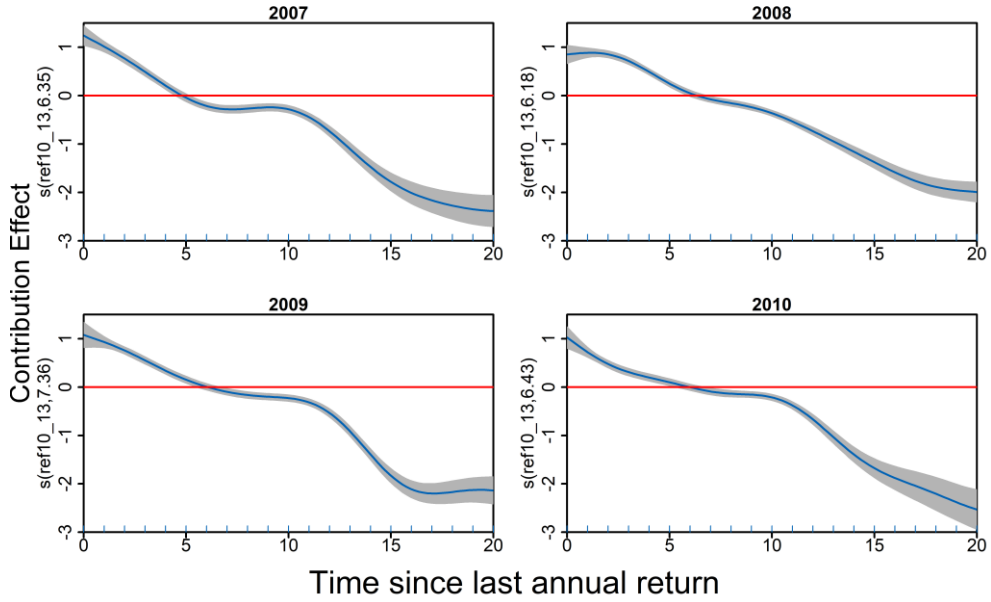


Figure 5-13 GAM - Time since last annual return

- Total assets

From rug plots in Figure 5-14, the majority of observations is detected from 0 to 1,500,000, and it lacks of data towards the end. The confidence limits are relatively narrow compared to the latter observations.

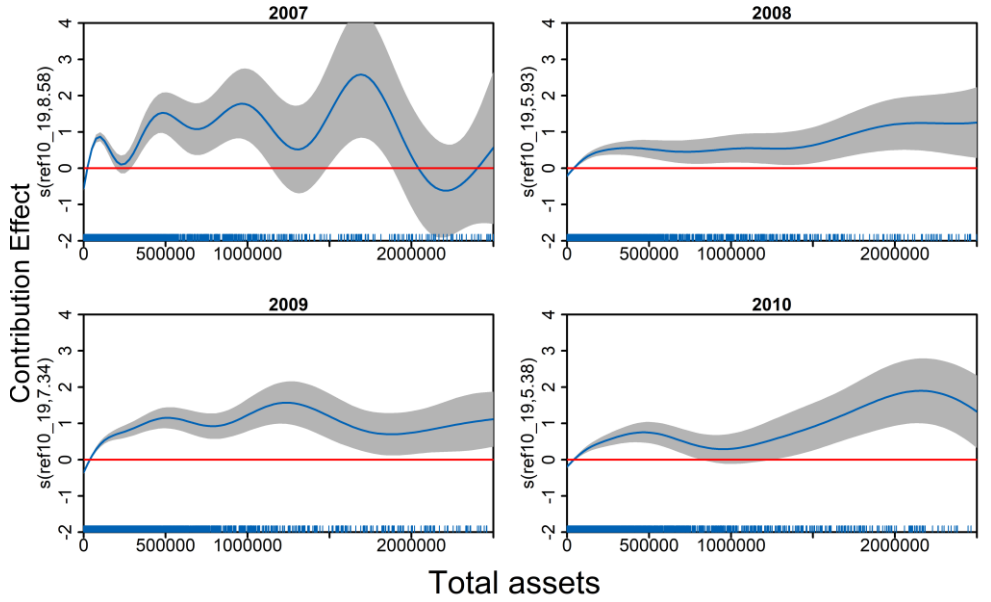


Figure 5-14 GAM - Total assets

Smooth functions in 2007 and 2009 are fluctuate more than in other years, which is therefore difficult to explain in detail. The smooth functions can be divided into

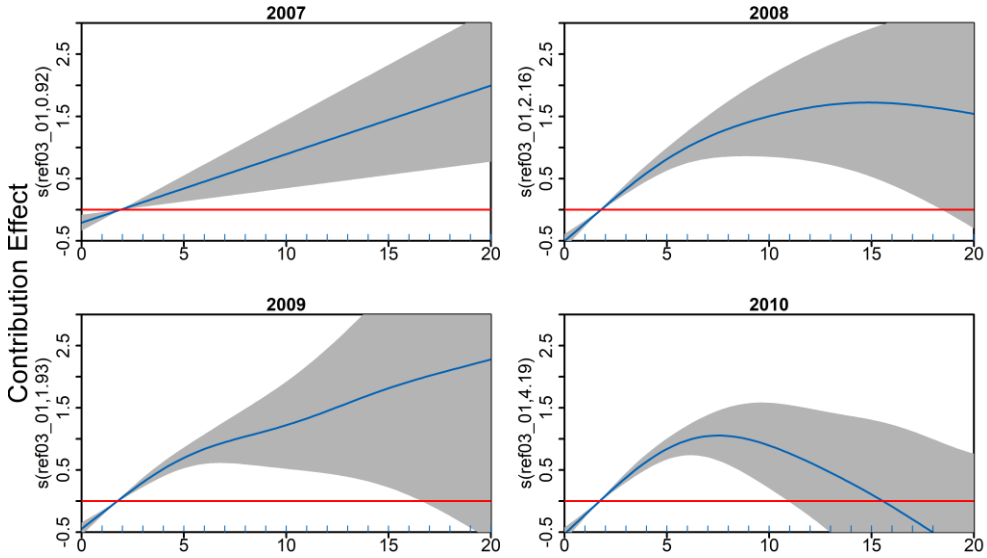
several parts to discuss. From 0 to 500,000, a significantly positive trend with relatively narrow confidence limits is observed. Although several waves are observed in 2007, its trend goes upward in general. From 500,000 to 1,000,000 decreases trend is observed except in 2008, and a confidence band becomes larger. From 2,000,000 to 2,500,000 a sharply increasing trend is observed mainly in 2007 with a reasonable confidence band. However, extremely wide confidence band in 2009 makes start-ups' performance less predictable. Above 2,500,000, influence is sensitive to the business cycle, with a wide confidence band. Only the smooth curve in 2009 observes a rising trend. Hence, the start-ups' performance is less predictable for large total assets.

In summary, increasing the amount of asset help SME's development. Yet, unlimited financial support would not work especially during the peak of credit crisis. Besides, when at the peak of credit crisis, its impact on this variable is obvious especially on the large value.

5.2.4.2 Non-start-ups

For non-start-ups, all the 7 variables are statistically significant (Table 5-13, p-value <0.05) as well, and thus important in the prediction model. Of the 7 independent variables, Pp worst (company DBT - industry DBT) in the last 12 months, Time since last derogatory data item (months), Lateness of accounts, and Time since last annual return show the greatest influential behaviour between 2007 and 2010 (smallest values of p-value in Table 5-13). Most of the predictor variables exhibit a significant non-linear effect for modelling (EDF >1), No. of 'current' directors in 2007 and 2008, Total value of judgements in the last 12

months in 2009 (EDF ≈ 1). In the following sections, smooth function plots



(Figure 5-15 - Figure 5-21) are discussed in detail.

Table 5-13 Effective degrees of freedom and approximate significance of each GAM smoothed term of non-start-ups

Smoothed term	EDF				p-value			
	2007	2008	2009	2010	2007	2008	2009	2010
No. of 'current' directors	0.91	1.07	1.93	4.18	9.00×10^{-4}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Pp worst (company DBT - industry DBT) in the last 12 months	3.89	2.9	3.91	5.13	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Total value of judgements in the last 12 months	2.37	2.99	0.93	3.08	6.91×10^{-10}	8.46×10^{-5}	1.67×10^{-4}	2.64×10^{-5}
Time since last derogatory data item (months)	6.26	7.23	7.67	7.32	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Lateness of accounts	5.94	7.35	4.88	7.03	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Time since last annual return	4.25	5.35	6.53	6.1	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}
Total fixed assets as a percentage of total assets	3.82	7.31	14.8	7.21	6.54×10^{-9}	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}

Notes: 2008 model uses a small ridge penalty added to the smoothing penalty so that the whole term

- No. of 'current' directors

Although non-linear relation only observed in 2009 and 2010, given the wide confidence bands, the suitable observed ranges are linear (Figure 5-15). Hence, a clearly positive trend is observed. This is a very strong implication suggesting

that the larger the director group, the more non-start-ups' probability of being 'good'. However, during the recovery period, this probability may reduce while increasing the number of current directors.

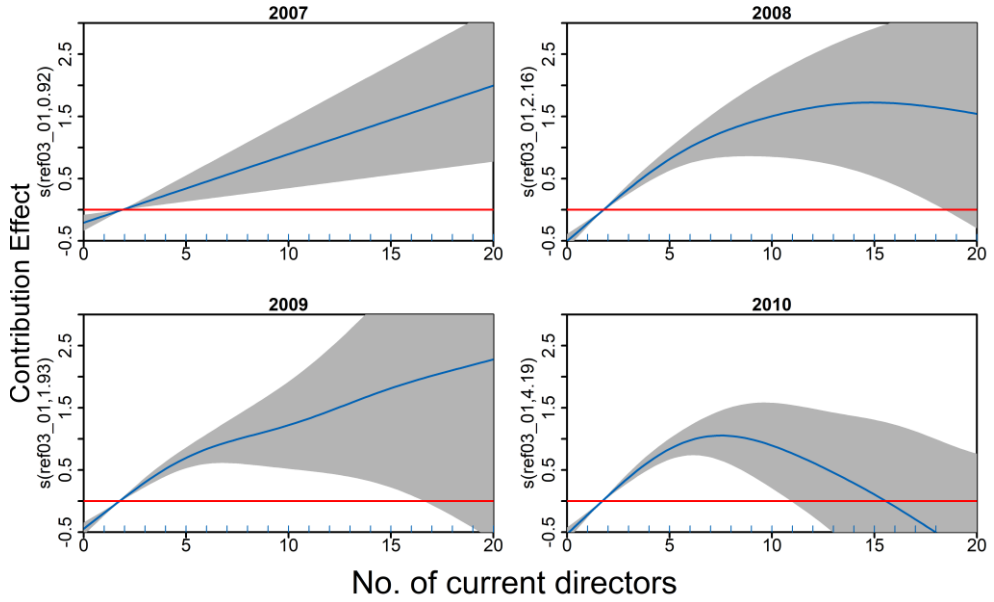


Figure 5-15 GAM - No. of 'current' directors

- worst (company DBT - industry DBT) in the last 12 months

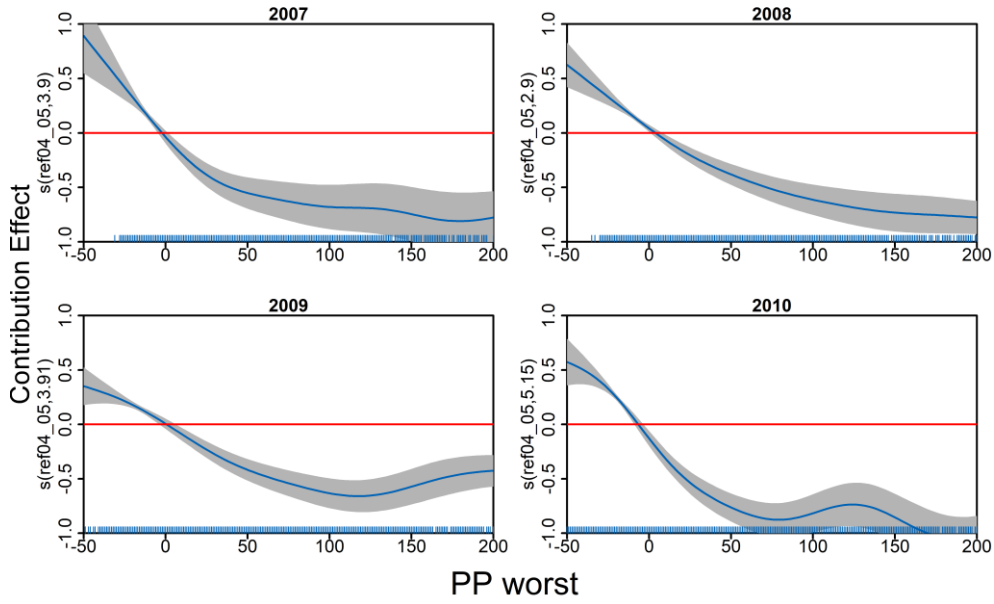


Figure 5-16 GAM - PP worst (company DBT - industry DBT) in the last 12 months

DBT refers to ‘Days beyond Terms’, which shows how rapidly firms transact their liabilities. PP worst (company DBT - industry DBT) in the last 12 months compares the company’s performance with that of the corresponding industry. Its smooth function can be divided into two parts to analysis (Figure 5-16). The first part is below 150. There is a negative trend consistently over the four years. It can conclude that the shorter non-start-ups pay invoice back, the lower their credit risk. The second part is above 150. This variable switches trends over time. Large value makes non-start-ups much more sensitive during business cycle especially.

- Total value of judgements in the last 12 months

A decreasing linear trend is only observed in 2009 (Figure 5-17). Majority observations range from 0 to 50,000, and during this range, a clear negative trend is observed with relatively narrow confidence band. Above this range, confidence limits expand since there is a lack of observations or outliers.

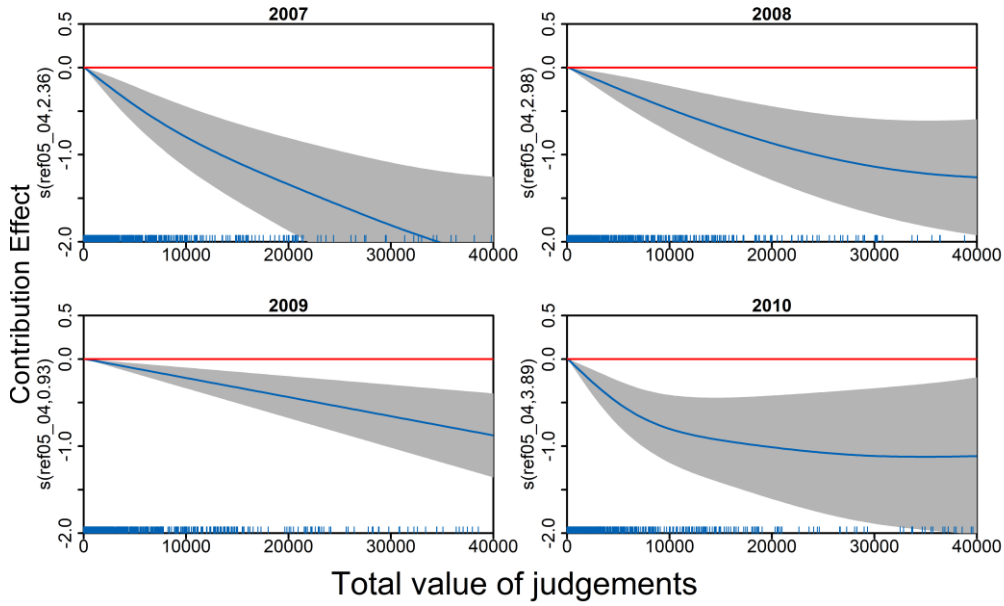


Figure 5-17 GAM - Total value of judgements in the last 12 months

- Time since last derogatory data item (months)

The smooth function plot (Figure 5-18) can be divided into three parts for discussion. The first part is from beginning to the first peak of the curve. Generally, a positive influence with fluctuations is presented from 2007 to 2010. This impact is very significant because of a very narrow confidence band. The second part is from the first peak to the 150th month. The curve becomes unstable, and it is

difficult to explain the trend. The third part is after the 150th month. The curve still increases in 2007 and 2008 but decreases in 2009 and 2010 with an acceptable width confidence limit. Changes on business cycle change the directions of these variables.

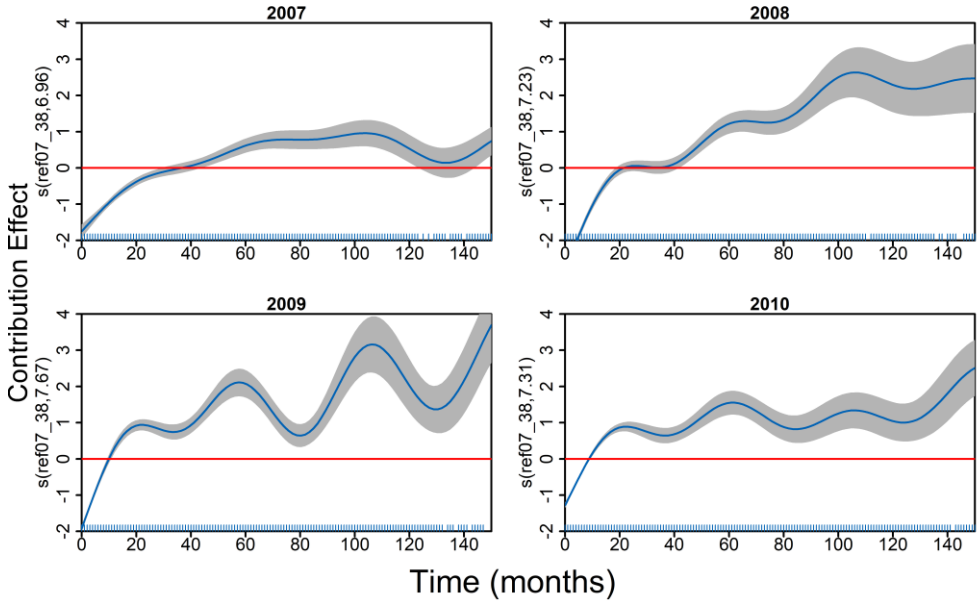


Figure 5-18 GAM - Time since last derogatory data item (months)

Similarly, the derogatory data is especially important if the record is more recent (from 0 to around 25 months). Exceeding the 25th month, it is difficult to find its pattern since the curve is much more fluctuate.

- Lateness of accounts

It is clear from Figure 5-19 that the changes in the economic environment have great impact on this variable due to the different shape of plots. During the normal period, and the beginning of the credit crisis, the smooth function curve is relatively flat, and around zero. After 50, both curves rise gradually with much wider confidence limits. However, during the peak of the credit crisis and recovery period, the curve shows a clear negative trend with tiny confidence limit, after that rising trend is observed but with boarder confidence bands.

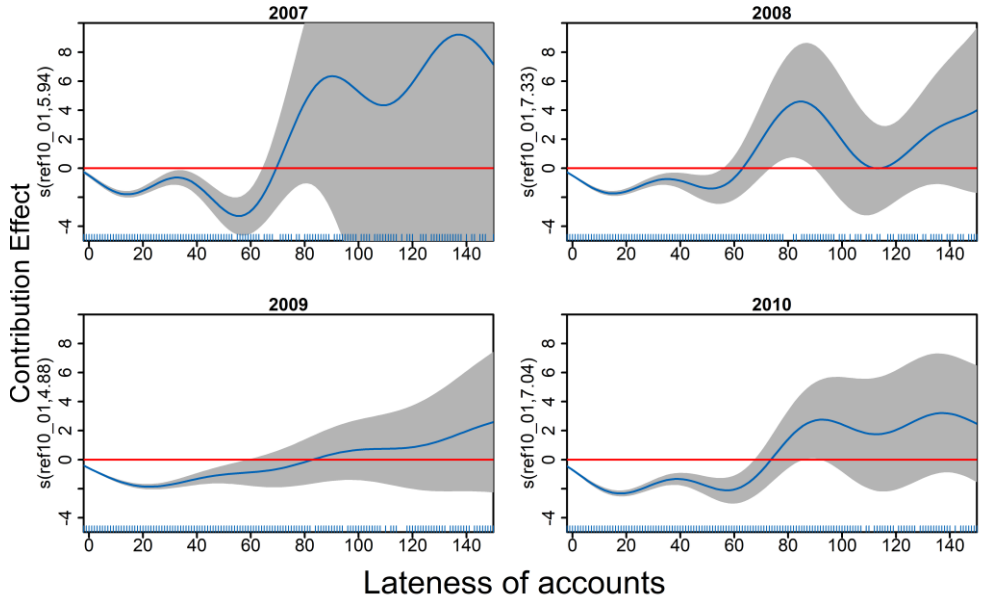


Figure 5-19 GAM - Lateness of accounts

- Time since last annual return

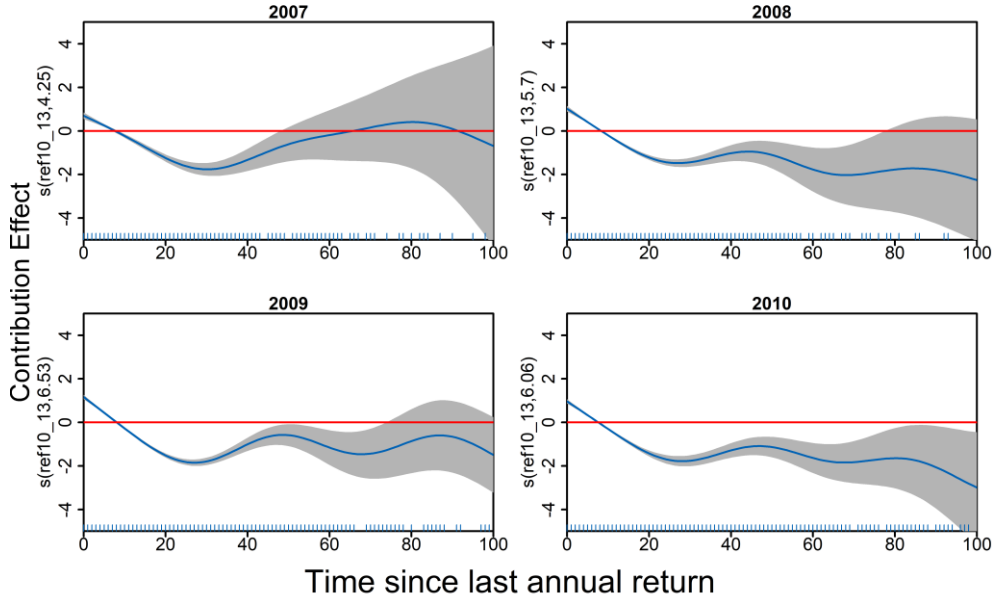


Figure 5-20 GAM - Time since last annual return

From rug plot in (Figure 5-20), a significant number of observations locate below 100. An initially decreasing trend is observed approximately from 0 to 25. Confidence limits of this range are narrow, and therefore, this is a very strong

conclusion. As the time increase, the confidence limit becomes wider gradually. After that, there is an increasing trend observed in the range from 25 to 50.

- Total fixed assets as a percentage of total assets

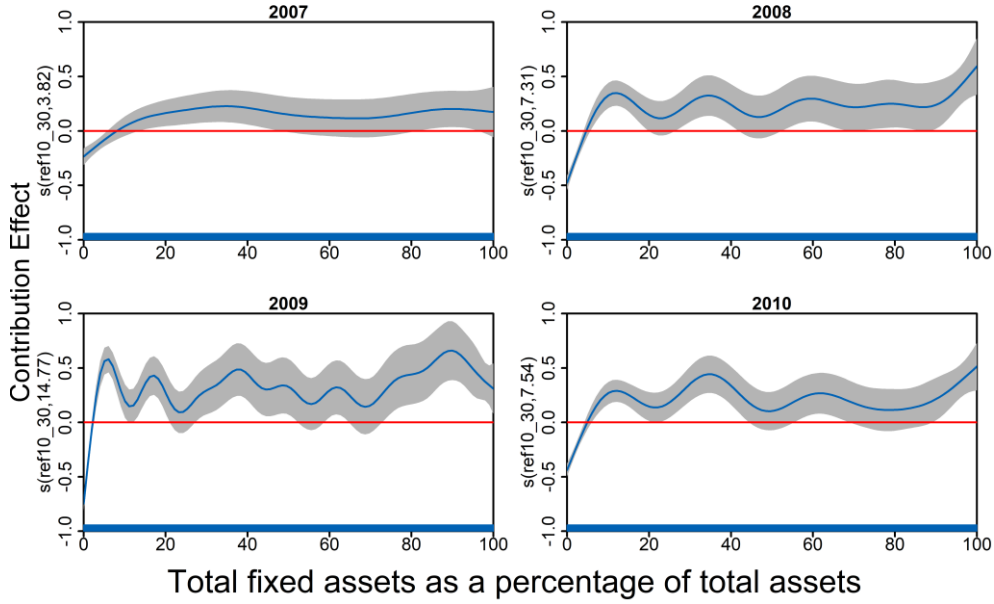


Figure 5-21 GAM - Total fixed assets as a percentage of total assets

As shown in Figure 5-21, the initial contribution effect falls year by year with the outbreak of the crisis, and it begins at approximately -0.5 in 2007, 2008 and 2010 but there is a sharp drop to -1 in 2009. observations range from 0 % to 100%, and the confidence limits are relative narrow below 20%. Initially, from 0% to 20%, a sharply rising trend is observed, especially in 2009. After that, the smooth functions in 2009 are fluctuates more and unstable because only a large EDF can describe the trend, which is therefore difficult to explain in detail. Yet, the smooth function remains an acceptable range with wide confidence limits. Hence, the non-start-ups' performance is less predictable for a large proportion of fixed assets. In summary, increasing the proportion of fixed assets help SME's development.

5.2.4.3 Summary of GAMs

In this section, GAM is applied to analyse imputed data. Changes in the economic environment have an influence on the movement of the relationship between dependent and independent variables.

Regarding directors' information, performance of both start-ups and non-start-ups have similarity. GAM model concluded that enlarging the size of board would help in 2009 to some extent. Specially, an increase of number of directors, directors aging between 40 and 60, and at least one director holding shares could good for start-ups survival in 2009. During the credit crisis, the original board of directors failed to achieve their responsibilities, and this is considered as a failures and weakness of corporate governance, which resulted in financial crisis in 2009 (Kirkpatrick, 2009). In addition, Aebi, et al. (2012) pointed out bank's crisis performance do not enhanced by standard corporate governance mechanisms. This means that exceptional times call for exceptional measures. This could be partly explained the reasons why it is necessary to expand the board of directors in 2009. While in the normal periods, Payne, Benson, & Finegold (2009) suggested that a long-term tenure improves the quality of the board and financial performance because it is associated with greater experience, commitment, and knowledge about the firm and its business environment.

In terms of the accounting information, an increase of assets for start-ups, and fixed assets for non-start-ups help to survive during the credit crisis but it does not mean it can increase unlimitedly.

Finally, there is a comparison of previous relevant credit history in both start-ups and non-start-ups since they use the same set of variables in the following. Total value of judgements in the last 12 months (Figure 5-10 and Figure 5-17): For both start-ups and non-start-ups, an initially negative trend of smooth function can be observed. Yet, increasing trend is observed in 2009 and 2010 for start-ups, and in 2007, 2008, and 2010 for non-start-ups with wide confidence limits, which is out of the suitable observed range. During the peak of credit crisis, the change on this variable is more sensitive for non-start-ups because of different slopes.

Time since last derogatory data item (months) (Figure 5-11 and Figure 5-18): For both start-ups and non-start-ups, the shape of smooth function curve is similar through 2007 to 2010 with a sharply increasing trend at initial. Since separation of

start-ups and non-start-ups bases on time of establishment, the maximum value of this variable is less than 30 for start-ups and over 150 for non-start-ups. Therefore, the overall trend for start-ups is increasing and relative flat although a wide confidence band is observed at large value. Regarding non-start-ups, smooth functions have more fluctuations, especially comparing the curve in 2009 and 2010. In summary, at the peak of the credit crisis, the smooth curve of start-ups is monotonously increasing, which means the longer the time, the lower the credit risk, while it becomes less intuitive for non-start-ups. At the normal period, it can conclude for both start-ups and non-start-ups that the shorter the time, the larger the credit risk although there are some small fluctuations.

Lateness of accounts (Figure 5-12 and Figure 5-19): Likewise, time effect makes a huge different of range of this variables for start-ups and non-start-ups. Confidence limits are extremely wide of large value for non-start-ups. For both start-ups and non-start-ups, an overall negative trend is found when value of the variable is negative. However, change on economic environment is significant only for non-start-ups.

Time since last annual return (Figure 5-13 and Figure 5-20): Confidence band expands when time exceeds 15 for start-ups and 50 for non-start-ups, and once the value is larger than 100, it becomes useless since the confidence area includes zero. In summary, as the time increases, the confidence limits become wider, and it shows a clearly negative trend since recent months, and then becomes flat or volatile.

5.2.5 Model Performance of Cross-section Analysis

This section provides the result of prediction models: logistic regression, shrinkage regression, and GAMs. Regarding coefficients, it is surprising that coefficient of logistic regression using stacked data shrinks to near zero, which has a similar effect as shrinkage regression. Besides, coefficients are unstable over the years, and therefore, a single model would not fit the prediction of SMEs performance over the years. GAMs results and unstable coefficient of logistic regression and

shrinkage regression prove that it is necessary to further explore the impact due to the change on economy.

Table 5-14 provides a comparison of AUROC among four models. Logistic regression with weights using stacked data always provides a better prediction performance although there is no clear difference among the models. It is surprising that the performance of lasso or ridge regression does not exceed that of logistic regression given using the same dataset with WoE transformation.

Table 5-14 AUROC on the test sample

Models	Start-ups				Non-start-ups			
	2007	2008	2009	2010	2007	2008	2009	2010
Stacked LR	0.866	0.882	0.857	0.877	0.808	0.881	0.886	0.857
LR WoE	0.831	0.848	0.858	0.823	0.769	0.837	0.888	0.824
Lasso WoE	0.826	0.845	0.855	0.821	0.765	0.830	0.885	0.820
Ridge WoE	0.825	0.845	0.857	0.820	0.763	0.836	0.888	0.818

There are contradictions regarding the significance of variable. Lasso regression can screen important variables, and it shows that proportion of current directors to previous directors in the last year should be removed from the model in 2008 and 2009, total value of judgements in the last 12 months should be discarded from the model in 2008 for start-ups SMEs. Yet, the result of logistic regression with WoE transformation shows that proportion of current directors to previous directors in the last year is significant in both 2008 and 2009, while total value of judgements in the last 12 months totally loses its significance. The change in the sign of the variable proportion of current directors to previous directors in the last year is found on logistic regression with stacked data and MICE procedure. Total value of judgements in the last 12 months is significant in 2008 and not in 2009 in the logistic regression with stacked data model.

Regarding non-start-up SMEs, results from lasso regression also show that except region in 2009, others need to be removed, total value of judgements in the last 12 months in 2009 as well. However, result of logistic regression with WoE

transformation and logistic regression with stacked data shows that these two variables are significant at 1% (region in logistic regression with stacked data model codes as dummy variables, hence it is difficult to judge its significance), which means they have a reliable prediction power, and should remain in the model.

5.3 Panel Models

The previous section shows the analysis result based on the cross-section for start-ups and non-start-ups, and the goodness-of-fit is satisfactory. Yet, it is reasonable to believe that the significant change in the macroeconomic environment affected the SMEs' performance during the credit crisis. Cross-section analysis based on firm-specific variables assumes that SMEs' performance is not related to the macroeconomic environment. Additionally, because of the large variation observed in the coefficient estimated in the last section, it is possible that changes in the distribution of probabilities are due to change in the macroeconomic variables.

To capture the time effect, the single period logistic regression model is developed to a multi-period logistic regression model. In the following, logistic regression with panel data only use firm-specific variables is initially used. After that, analysis of adding annual dummy variable and MVs to solve the problem of time effect during the observed period.

Coefficients and standard errors for start-ups and non-start-ups are shown in the Table 5-18 and Table 5-19 respectively, while goodness-of-fit of those models for training samples and testing samples are given in Table 5-20.

5.3.1 Macroeconomic Variables (MVs)

The drawback of annual dummy variables can be overcome by using MVs since MVs are able to capture the market movement and provides accountable results of how MVs affect SMEs' performance during the credit crisis.

In order to explore the time effect of SMEs performance during credit crisis, potential MVs (Table 5-15) should remain consistent during the period and have strong connection to SMEs performance in the UK. When introducing MVs, correlation among variables needs to be considered carefully, see Table 5-16. A high correlation between MVs is a potential problem, since this could lead to multi-collinearity within the PD model and therefore distort the parameter estimates (Bellotti and Crook, 2012).

Table 5-15 UK Macroeconomic data from 2007 to 2010

	2007	2008	2009	2010
GDP growth rate (%)	2.4	-0.5	-4.2	1.7
Unemployment rate (%)	5.3	5.7	7.6	7.9
Consumer price inflation (%)	2.3	3.6	2.2	3.3
FTSE-100 Index: % change	3.8	-31.3	22.1	9.0
FTSE-All-Share Index: % change	2.0	-32.8	25	10.90
Interest rate (%)	5.55	5.09	2.21	2.8

Both Akaike information criterion (AIC) (Akaike, 1987) and Bayesian information criterion (BIC) are suitable measures to compare maximum likelihood models. Given two models fit on the same data, the model with the smaller value of the information criterion is considered to be better. They are defined as:

$$\mathrm{AIC} = -2 \times \log(\mathrm{likelihood}) + 2 \times \mathrm{K} \tag{44}$$

$$\mathrm{BIC} = -2 \times \log(\mathrm{likelihood}) + \log(\mathrm{N}) \times \mathrm{K} \tag{45}$$

where: k = model degrees of freedom, N = number of observations

Table 5-16 Correlation of UK Macroeconomic data

	GDP growth rate (%)	Unemployment rate (%)	Consumer price inflation (%)	FTSE-100 Index: % change	FTSE-All-Share Index: % change	Interest rate (%)
GDP growth rate (%)	1					
Unemployment rate (%)	-0.43	1				
Consumer price inflation (%)	0.29	-0.01	1			
FTSE-100 Index: % change	-0.26	**0.63**	**-0.74**	1		
FTSE-All-Share Index: % change	-0.29	**0.68**	**-0.71**	**0.99**	1	
Interest rate (%)	0.60	**-0.97**	0.17	-0.69	-0.74	1

However, there is an argument of which one should be used (D. Anderson and Burnham, 2004; Yang, 2005). AIC is susceptible to over-fitting the data, whereas BIC is susceptible to under-fitting the data when the goal is to maximize predictive discrimination. The reason is that they penalize the free parameters differently, i.e., 2K in AIC, K*log(N) in BIC. Considering this, both AIC and BIC are provided for discussion.

Initially, AIC and BIC from logistic regression with firm-specific model is shown as a benchmark when considering adding time effects. After that, selected MVs will add to the model individually. Details can be found in Table 5-17, and each MV is highly statistically significant (P-value=0) and the coefficients have no obvious anomalous signs.

Figlewski, et al. (2012) introduced three aspects to describe the macroeconomic environment: general macroeconomic conditions, the direction of the economy, and financial market conditions. In this research, consumer price inflation (CPI) and unemployment rate are the potential variables for general macroeconomic. Gross domestic product (GDP) growth rate is the variable for direction of the

economy. FTSE index and interest rate are the potential variables for financial market conditions.

The common perception is that high unemployment rate is bad for the economy so that they are not good for SMEs' performance. In other words, it is expected that there is an adverse relationship between 'good' SMEs and employment rate, and this can be confirmed in this research, see Table 5-17.

Likewise, high CPI is not a good thing for SMEs' development as inflation can boost the worthiness of their business, increase their cost and decrease customer's buying power. Yet, from the perspective of a firm whose outstanding debt is in nominal dollars, inflation reduces the real value of its required debt service payments, which might make it less likely to default (Figlewski, et al., 2012). CPI's effect is different for start-ups and non-start-ups in this research: negative influence for start-ups and positive influence for non-start-ups. Start-ups suffer more from their increasing costs and loss of customers from the 'credit crunch', but non-start-ups as debtors gain from inflation because they repay creditors with money that are worth less in terms of purchasing power.

Unemployment rate has a high correlation to FTSE index and interest rate (Table 5-16), and it has a higher either AIC or BIC. Therefore, CPI is determined to be as one of the MVs in this research.

GDP growth rate as an important MV is a norm in many researches since if the economy is growing rapidly, it is clearly in better health than if it is stagnant or shrinking. For both segments, GDP growth rate constantly shows a positive correlation with the SMEs performance (Table 5-17). As GDP growth rate reflects the economic direction, the results mean that the SMEs performance is improved if the economy is strong. GDP growth rate has a low correlation to other potential MVs, and AIC and BIC improve slightly so that it is a feasible option in this research.

FTSE is a share index and it should has been an ideal indication to reflect financial market conditions. The two stock market return variables (FTSE-100 Index: %

change and FTSE-All-Share Index: % change) have a high correlation of 0.99. Additionally, they are highly correlated to CPI and unemployment rate. Therefore, FTSE index variables are replaced by interest rate, but a high correlation of -0.97 was found for the unemployment rate and a moderate correlation of 0.6 to GDP growth rate.

Table 5-17 Analysis of individual MVs

	Start-up				Non-start-up			
	Sign	P-value	AIC	BIC	Sign	P-Value	AIC	BIC
Panel LR			80501.87	80766.48			66321.13	66586.67
Year dummy			79306.59	79600.6			66026.69	66321.74
GDP growth rate	+	0.0	80475.36	80749.77	+	0.0	66026.53	66301.91
UR	-	0.0	80311.54	80585.95	-	0.0	66317.41	66592.8
CPI	-	0.0	79669.63	79944.04	+	0.0	66226.3	66501.68
FTSE_ALL	+	0.0	80322.85	80597.26	-	0.0	66296.06	66571.45
Interest Rate	+	0.0	80399.14	80673.55	+	0.0	66271.12	66546.5
GDP+UR+FTSE			79306.59	79600.6				
GDP+CPI+FTSE			79306.59	79600.6				
GDP+CPI+IR			79306.59	79600.6			66026.69	66321.74

Notes: This table provides MVs sign, P-value, AIC and BIC for both segments.

UR: unemployment rate; CPI: Consumer price inflation; FTSE_ALL: FTSE-All-Share Index: % change

Other things equal, one might expect that high interest rates would also correspond to general tightness in the economy and increased difficulty in raising cash to make debt service payments (Figlewski, et al., 2012). A sharp decrease was found in 2009 where it was the period of credit crunch. More money went into the market so that stimulate consumption, and SMEs borrowed money at a lower interest rate.

To sum up, GDP growth rate, consumer price inflation, and interest rate are reasonable to be chosen to be MVs added into the panel model. CPI has the most influence on start-ups and GDP growth rate for non-start-ups.

5.3.2 Explanatory Models

For both start-ups and non-start-ups, their estimated coefficient and standard error of panel models are displayed in Table 5-18 and Table 5-19 correspondingly, and the goodness-of-fit result is shown in Table 5-20.

For the dummy variable model, the pre-crisis period is set as the reference category. The majority of time dummies significant throughout the 'credit crunch' for both segments. Even if one sets up a panel data model with firm-specific variables only, the results could not be misleading. The only exception is found in 2010 for non-start-ups. This implies that the recovery ability of matured firms is much better than newly established firms as the difference between the reference category and 2010 is not significant. After the crisis, the non-start-ups can recover to the pre-crisis performance soon.

In terms of MVs models of start-ups, changes in coefficients among different models are small. Additionally, MVs are significant at 0.1% and the sign of coefficients are consistent with that of individual MV analysis when adding all selected MVs.

Concerning non-start-ups MVs models, coefficients are similar for firm-specific variables to other models. Changes in coefficients among different models are small. A negative relationship between SMEs' performance and interest rate is found, but when adding only interest rate to the model their relationship is positive. Additionally, CPI and interest rate are no longer significantly, but they are significant at 0.1% in individual MV analysis. The reason of variables are significant in single variable analysis and insignificant in multiple regression analysis may be because the macro variables are fairly highly correlated with one another, coefficient estimates change substantially when all MVs are combined in a single model (Figlewski, et al., 2012). For the economy, an increase in interest rate will tend to lower the inflation and slower the economic growth so that increasing the cost of bank loans (cost of borrowing).

Regarding goodness-of-fits for the models, little improvement is found when adding annual dummy variables or MVs to control the influence of time effect for both start-ups and non-start-ups, although dummy variables and MVs are significant. This means there is a relationship between SMEs' performance and significant year dummy variables or MVs, but it does not help improve the ability to separate good/bad. Compared with the panel models, cross-section models using stacked imputed data outperforms the performance of AUROC except for non-start-ups in crisis-period. As logistic regression only considers the firm-specific difference, it implies that start-ups are established to be vulnerable with a higher default rate in general regardless of any economic conditions. Their classification models are robust in any condition, and their performance can be well explained by the specific corporate variables. For non-start-ups, matured firms are more subject to the great shock of MVs since they have more connection to the outside environment. With MVs, the goodness-of-fits of the model can be improved.

In summary, the panel data model with MVs is recommended without considering the access to data with multiple periods. The model is superior especially for non-start-up because the ability of classification is at least as the model with firm-specific variables only and exceeds for non-start-ups when there is a great shock on the economic environment.

Table 5-18 Start-up random effect panel data models parameter estimation

Variables	Firm-specific variables only		+ Dummy		+ GDP+CPI+IR	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Legal Form 1	-2.370***	-0.28	-2.459***	-0.28	-2.459***	-0.28
Legal Form 2	-2.439***	-0.09	-2.486***	-0.09	-2.486***	-0.09
Legal Form 3	-1.670***	-0.27	-1.713***	-0.27	-1.713***	-0.27
Legal Form 5	-1.483***	-0.15	-1.508***	-0.15	-1.508***	-0.15
Legal Form 6	0	(.)	0	(.)	0	(.)
Legal Form 7	-1.664***	-0.12	-1.702***	-0.12	-1.702***	-0.12
Legal Form 8	0	(.)	0	(.)	0	(.)
Legal Form 9	0	(.)	0	(.)	0	(.)
Region South East	0	(.)	0	(.)	0	(.)
Region South West	0.09	-0.06	0.08	-0.06	0.08	-0.06
Region North East	-0.145***	-0.03	-0.163***	-0.03	-0.163***	-0.03
Region North West	-0.08	-0.07	-0.1	-0.07	-0.1	-0.07
Region East Midlands	-0.715***	-0.15	-0.821***	-0.15	-0.821***	-0.15
Region West Midlands	-0.237***	-0.04	-0.239***	-0.04	-0.239***	-0.04
Region East England	0.26	-0.92	0.32	-0.91	0.32	-0.91
Region Yorkshire	-0.0876*	-0.04	-0.118**	-0.04	-0.118**	-0.04
Region Scotland	0.08	-0.05	0.08	-0.05	0.08	-0.05
Region Wales	0.334***	-0.05	0.322***	-0.05	0.322***	-0.05
Region North Ireland	0.07	-0.04	0.03	-0.04	0.03	-0.04
Region Others	-0.161*	-0.07	-0.162*	-0.07	-0.162*	-0.07
Region South East	-0.124**	-0.04	-0.155***	-0.04	-0.155***	-0.04
Proportion of current directors to previous directors in the last year	0.0596***	-0.01	0.0510***	-0.01	0.0510***	-0.01
Oldest age of current directors/proprietors supplied (years)	0.0147***	0	0.0145***	0	0.0145***	0
Number of directors holding shares	0.747***	-0.02	0.782***	-0.02	0.782***	-0.02
Total value of judgements in the last 12 months	-0.0000142*	0	0.0000158*	0	0.0000158*	0
Time since last derogatory data item (months)	0.360***	0	0.381***	0	0.381***	0
Lateness of accounts	-0.170***	0	-0.171***	0	-0.171***	0
Time since last annual return	-0.142***	0	-0.140***	0	-0.140***	0
Total assets	.000000980*	0	.00000103*	0	.00000103*	0
2008.year			-0.859***	-0.03		
2009.year			-0.469***	-0.03		
2010.year			-0.784***	-0.03		
GDP_rate					4.014***	-0.46
CPI					-54.35***	-1.67
IR					7.736***	-0.82
_cons	1.417***	-0.11	1.953***	-0.11	2.678***	-0.12

* p<0.05, **p<0.01, *** p<0.001

Table 5-19 Non-start-up random effect panel data models parameter estimation

Variables	Firm-specific variables only		+ Dummy		+ GDP+CPI+IR	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Legal Form 1	-6.584***	-0.34	-6.420***	-0.34	-6.420***	-0.34
Legal Form 2	-6.418***	-0.22	-6.239***	-0.22	-6.239***	-0.22
Legal Form 3	-5.718***	-0.48	-5.549***	-0.48	-5.549***	-0.48
Legal Form 5	-6.196***	-0.26	-5.972***	-0.25	-5.972***	-0.25
Legal Form 6	0	(.)	0	(.)	0	(.)
Legal Form 7	-6.471***	-0.24	-6.296***	-0.24	-6.296***	-0.24
Legal Form 8	11.34***	-1.14	11.03***	-1.12	11.03***	-1.12
Legal Form 9	-6.954***	-1.17	-6.693***	-1.19	-6.693***	-1.19
Region South East	0.384***	-0.06	0.387***	-0.06	0.387***	-0.06
Region South West	0.116**	-0.04	0.119**	-0.04	0.119**	-0.04
Region North East	0.208**	-0.07	0.207**	-0.07	0.207**	-0.07
Region North West	-4.421***	-0.24	-4.288***	-0.24	-4.288***	-0.24
Region East Midlands	0.06	-0.04	0.06	-0.04	0.06	-0.04
Region West Midlands	0	(.)	0	(.)	0	(.)
Region East England	-0.122**	-0.04	-0.0904*	-0.04	-0.0904*	-0.04
Region Yorkshire	0.133**	-0.05	0.138**	-0.05	0.138**	-0.05
Region Scotland	0.481***	-0.06	0.472***	-0.06	0.472***	-0.06
Region Wales	0.06	-0.05	0.08	-0.05	0.08	-0.05
Region North Ireland	0.253***	-0.07	0.269***	-0.07	0.269***	-0.07
Region Others	-0.02	-0.05	-0.01	-0.05	-0.01	-0.05
No. Of 'current' directors	0.263***	-0.01	0.257***	-0.01	0.257***	-0.01
Pp worst (company DBT - industry DBT) in the last 12 months	-0.00598***	0	-0.00595***	0	-0.00595***	0
Total value of judgements in the last 12 months	-0.0000302***	0	-0.0000302***	0	-0.0000302***	0
Time since last derogatory data item (months)	0.0518***	0	0.0502***	0	0.0502***	0
Lateness of accounts	-0.0316***	0	-0.0314***	0	-0.0314***	0
Time since last annual return	-0.0504***	0	-0.0487***	0	-0.0487***	0
Total fixed assets as a percentage of total assets	0.00853***	0	0.00846***	0	0.00846***	0
2008.year			-0.186***	-0.04		
2009.year			-0.401***	-0.04		
2010.year			0.008	-0.04		
GDP_rate					7.010***	-0.51
CPI					0.67	-2.08
IR					-1.85	-0.95
_cons	7.416***	-0.23	7.411***	-0.23	7.330***	-0.23

* p<0.05, **p<0.01, *** p<0.001

Table 5-20 AUROC of panel models

Models			2007	2008	2009	2010
Start-ups	Panel data	Train	0.861	0.877	0.851	0.875
		Test	0.867	0.885	0.852	0.879
	Year dummy	Train	0.862	0.877	0.851	0.875
		Test	0.868	0.885	0.852	0.88
	GDP+CPI +IR	Train	0.862	0.877	0.851	0.875
		Test	0.868	0.885	0.852	0.88
Non-start-ups	Panel data	Train	0.798	0.89	0.895	0.83
		Test	0.799	0.892	0.895	0.836
	Year Dummy	Train	0.798	0.89	0.895	0.83
		Test	0.799	0.892	0.895	0.836
	GDP	Train	0.799	0.89	0.895	0.83
		Test	0.8	0.892	0.895	0.836
	GDP+CPI +IR	Train	0.798	0.89	0.895	0.83
		Test	0.799	0.892	0.895	0.836

5.4 Summary

SMEs' failure prediction consists of separating the firms with a high probability of future failure from those that are non-default ('good'). This chapter initially gives an investigation of cross-section analysis. Considering a great number of missing data, two methods are applied to handle the problem: MICE and WoE. On the one hand, due to the application on MICE, there are multiple datasets. Two approaches are used to analyse theses dataset. One is to run logistic regression for multiple datasets and use Rubin's rule to combine the coefficient estimates and standard error. Another one is to stack entire all imputed dataset as a huge dataset, and then run a logistics regression with weights to fix the problem of underestimating standard errors. On the other hand, WoE is a favorite tool in credit scoring, and it is able to transform missing data to a specific value. Logistic regression has links with WoE so that it is used to model the data after WoE transformation. Shrinkage method can balance the bias and variance trade-off. It can help prevent the model from overfitting by adding penalty coefficient. Finally, GAM is used to explore the non-linear relationship among the variables. GAM is not easily interpreted if not control the EDF, in particular when they involve complex non-linear effects of some or all of the predictor variables although there are instances where generalized additive models may yield a better fit than generalized linear models, while parameter estimates can be used to predict or classify new cases simply and straightforwardly. It would be preferable to rely on a simple well-understood model for predicting future cases than on a complex model that is difficult to interpret and summarise. GAM model would be possible to apply on the industry as long as the EDF is controlled.

Cross-section analysis may be limited. Coefficient estimates and variable significance fluctuate among the models during the credit crisis. Obviously, ignoring the time effect and other latent effects cannot deeply explore the further relationship between dependent and independent variables although all of the models have good predictive performance. In addition, since changes on economic environment affect the movement of some variables, it is necessary to

investigate the truth behind that. Therefore, panel models are followed and try to control the time effect using year dummy variables and MVs. For start-ups, time effect does not help increase the ability to separate good from bad. Using dummy variables, it can be found that non-start-ups return to the pre-crisis period level in a short time. CPI and interest rate are insignificant, but panel models outperform cross-section models during the crisis period for non-start-ups.

CHAPTER 6 CONCLUSIONS AND DISCUSSIONS

SMEs are the research subject of this thesis because of their importance to a nation's development in numerous aspects, and less focus on SMEs study in credit risk modelling. Moreover, SMEs still have problems in receiving lending since the majority of them do not have publicly traded equity and certified audited financial statements leaving a problem of insufficient data for model building considering their characteristic (S.-M. Lin, 2007a). Yet, one positive fact is that Basel IV (Basel reforms) continues to adjust the capital requirement, while the issue of IFRS 9 for SMEs helps their financial information disclosure.

Preceding chapters have addressed the research questions by discussing how IFRS 9 influences banks and SMEs, how to deal with missing data with the imputation techniques, how to model default probability with imputation data and, and how to build credit risk models considering the time effect. This chapter provides a conclusion of this thesis by answering the research questions and points out the future possible research direction and limitations.

It is crucial for banks to manage the risks that they are exposed to, particularly the credit risk. This allows them to be timely anticipate any significant changes in the risk in order to allocate appropriate levels of provision. Credit risk management has experienced a material development in recent decades, especially since the implementation of the Basel Accord, and has become one of the most popular research interests in the financial sector. A correctly estimated probability of default is the core input factor for modelling and measurement of credit risk for the IRB approach and is a widely accepted strategy by financial institutions as well as the supervisory authorities globally. Yet, the estimates of the probability of default can be regarded as one of the critical challenges in credit risk management since accurate estimate might result in reasonable judgement of SMEs' performance. In fact, these were the reasons behind the emergence of the recent global financial crises as the misperception of default probabilities for mortgages and structured

credit products caused great stress on the financial system which had been extended through credit derivatives on global markets.

On the other hand, the new implement of IFRS 9 also provides a significant shock on modelling default risk. The IFRS 9 should improve credit risk management in banks in the long term, as well as internal processes of provision determination, reduce pro-cyclicality through recognition of timely credit losses and a higher level of transparency. All these measures will help improve financial stability in general.

As one of the most important goals, this thesis is looking for an appropriate and effective approach to develop SMEs credit risk model. Academic researchers, financial practitioners, and even regulators are paying growing attention to this topic. This research aims to investigate how SMEs perform using logistic regression, shrinkage regression, GAM models, and panel models with the potential predictors over the credit crisis. The findings from the analysis of results can assist financial institutions, especially banks in implementing the internal rating system to evaluate the performance of SMEs. To improve the financial institution's competitive advantage in the area of credit risk management, advanced methods and technologies must be developed to detect potential credit failures. The research provides empirical findings which give insight into credit rating and discussions for further study.

6.1 Research Questions Answered

In this section, the answer to the research questions will be provided.

- Except for WoE, is there another way to handle missing data better than the convention methods such as listwise deletion?

Listwise deletion has been the most prevalent method for missing data since it is uncomplicated to apply and available in majority data analysis software. If missing data is MCAR, the use of listwise deletion would not generate a significant problem if only a small number of missing data. Yet, this is not the situation for SMEs data. Generally, SMEs data is MAR with a higher volume of missing rate so that it is not

appropriate to use listwise deletion due to bias and a lack of statistical power. Listwise deletion used in the dataset with merely 1% complete observations meaningless to data analysis and models building in subsequent.

WoE has been proved its ability in the field of credit scoring with theoretical and empirical support. This thesis discusses an advanced alternative to deal with missing data MICE, which is uncommon in credit scoring. Under the assumption of MAR, MICE can generate multiple dataset and pool the coefficient estimates and standard error by Rubin's rules. The standard error produced by MICE would not be underestimated, which is a pitfall of the single imputation. Besides, for different types of variables, MICE can provide corresponding methods to impute the variables. Although listwise deletion and MICE methods cannot be directly compared, MICE would be more applicable to SMEs data.

- Does logistic regression using an alternative method provide an acceptable prediction accuracy for SMEs probability of default? In addition to logistic regression, is there a better approach to improve the prediction accuracy?

Logistic regression models have been industrial standard methods in credits scoring, and it is an ideal benchmark for other models' prediction accuracy. As known, the third step of MICE process is to estimate the coefficients for each imputed dataset by Logistic regression and combine them by Rubin's rules. However, it is criticized that it may cause overfitting and be complex to produce prediction accuracy. In light of this, an alternative is to logistic regression on the stacked imputed dataset with weights adjustments. To prevent overfitting, shrinkage regression models are used to build up credit risk modelling to trade off a slight rise in bias for a marked drop in prediction accuracy.

On the other hand, WoE as a prevalent method of data cleaning (dealing with missing data) in credit risk, is used to make a comparison using the same models with stacked data. The results show that the logistic regression model with stacked data predicts SME's performance accurately, even during the credit crisis.

- Is there any non-linear relation between independent variables and SMEs' performance? How to model with the non-linear effects?

Generalized Additive Models are statistical models in which the usual linear relationships between the response and predictor variables are replaced by non-linear 'smooths'. It has greater flexibility than traditional parametric modelling tools such as linear. It relaxes the usual parametric assumption and enables researchers to uncover structure in the (unobservable) relationship between the independent variables and the dependent variable. The GAM framework allows researchers to control the smoothness of the predictor functions to prevent overfitting. By controlling the fluctuations of the predictor functions, they can directly tackle the bias/variance trade-off. Essentially, the larger the number of estimated degrees of freedom, the wigglier the fitted model. Values of around 1 tend to be close to a linear term.

The marginal impacts on SMEs performance of some predictors do not increase (or decrease) linearly over the observed periods. The value of those predictors within a specific range will help withstand the crisis, such as directors ageing from 40 to 60 otherwise, decreasing their viability. In addition, the specific range will be time-varying. Take the variable Time since last derogatory data item (months) as an example, the smooth curve for start-ups is monotonously increasing in 2009, and this means the longer the time, the lower the credit risk. While the shorter the time, the larger the credit risk is in 2007. Besides, there are different marginal effects for start-ups and non-start-ups. For example, change in the economic situation would affect the lateness of accounts for non-start-ups but not for start-ups.

One of the issues related to GAMs is interpretability and flexibility. Linear models are easily understood, summarized, and communicated to others (e.g., in technical reports). Moreover, parameter estimates can be used to predict or classify new cases in a simple and straightforward manner. On the contrary, GAMs are not easily interpreted, considering how to control the fluctuations (EDF levels). Practitioners may struggle with the selection between an uncomplicated and well-

understood model and a complex poorly understood model that may have better prediction accuracy.

- During the credit crisis, there was a significant shock on macroeconomic, do these effects have a marked impact on the viability of SMEs? How to model the SME' performance due to the change in the macroeconomic environment?

To control the change in the macroeconomic environment, panel data models are introduced. The GDP growth rate, CPI and interest rate are selected as macroeconomic variables (MVs) to control the time effect. Regarding the individual MVs analysis for start-ups, a negative relationship between CPI and SMEs' performance is found while other MVs show a positive relationship. For non-start-ups, all MVs presents a positive relationship to the SMEs performance.

The prediction accuracy for start-ups is similar for cross-section analysis (Logistic regression) and panel data models with all selected MVs. That is, the sharp change in economic conditions have little impact on the viability of start-ups. This is not the case for non-start-ups. The AUROC outperforms for panel data models than that for cross-section models in 2008 and 2009. This means a great shock on MVs does impact on the performance of non-start-ups.

- How the implementation of IFRS 9 affects banks and SMEs?

The incurred loss provisioning model (IAS 39) is criticised for its backwards-looking characteristic as it only considers the past events and current conditions. This may be delayed recognition of credit losses and may postpone the provisioning preparing.

After redesigning the accounting standards, ECL (IFRS 9) was proposed to replace the incurred loss model. ECL are to be measured either on a (forward-looking) 12-month or a lifetime basis, depending on whether there has been a material increase in credit risk since initial recognition.

A key element of IFRS 9 is a forward-looking expected loss impairment model. IFRS 9 is not prescriptive about how exactly the changes in the credit/macroeconomic environment should be reflected in the expected loss estimation, but the standard requires that the expected loss estimate be forward-looking and incorporate available information at the time of estimation. That is, forecast information must be used in measuring ECL. Macroeconomic panel data modelling satisfies the requirements of measuring ECL and well suited to capture the credit behaviour and bring macroeconomic factors explicitly.

The ECL from IFRS 9 has a significant impact on banks and SMEs. Under the impairment approach in IFRS 9, the implement of ECL will lead to banks making higher provisioning as every loan to SMEs will bear small provisions from the day of origination, while loans that have had significant reductions in credit quality will incur larger provisions. This may reduce a bank's capital and profitability and cause greater provisioning volatility as the ECL forecasting horizon switches between 12 months and lifetime. However, it is expected to have a strong influence on banks' overall stability over the long- term period. For SMEs, especially those with poor or even empty credit history, lending from banks may be at a higher cost because of its high risk-weight, and the long-term loan may be less possible as banks have to take more provisioning. For this reason, IASB proposed an IFRS for SMEs to ease the financial reporting burden as well as reduce the cost of financial information disclosure.

6.2 Limitation and Suggestion for Further Research

Models are indispensable in quantifying and managing credit risk. Models rely on a range of data input based on a combination of historical data and risk assumptions and are critical in estimating the probability of default. However, models rely on the accuracy of inputs, and errors give rise to model risk. Model risk can range from errors in inputs and assumptions to errors in implementing or incorrectly interpreting a model and can result in significant losses to participants. Credit risk management poses specific challenges and limitation for quantitative modelling, but the problems can be considered for further research. A part of

observation was recorded in a stressed period, and one advantage is that risk would not be underestimated due to the data input reason (but would still be underestimated because of models' errors).

6.2.1 Dealing with Missing Data

From the dataset used in the thesis, one should have expected to look at the performance of SMEs, especially during the period of high volatility economically, but the reality may be covered by the missing data. It is more likely to problematically impute the dataset, even the use of MICE in such a recession period. There is no solution to compare true data and imputed data.

However, what can be done is to tune the parameters or apply more advanced methods to recover the true data as close as possible. Regarding MICE process, several parameters may be available to adjust to create more accurate estimates. There should be room for improvement through adjustments, such as the interaction between variables, variables transformation and the number of imputations. Although differences between the observed and imputed values do not mean a problem in the analysis. Besides handling missing data, MICE will have an effect on the selection of predictors which may be a concern as well.

Besides, considering the high missing rate in financial data, modelling credit risk for SMEs with non-financial data would be an alternative approach to lower the missing rate in the dataset. Edward I Altman, et al. (2008) showed that the use of non-financial variable can increase the prediction accuracy. Furthermore, non-accounting information can be updated frequently allowing financial institutions to correct their credit decisions in a timely manner.

As mentioned, the ultimate method for dealing with missing data is no missing data. Financial institutions have raised awareness of the risks involved with data issues and been able to identify ways to protect one of their most valuable resources, their data since the presence of the three pillars from Basel II With the

implementation of the Basel III and IFRS 9, the problem will be gradually improved in the future.

6.2.2 Data-quality Reject Inference

Except for the missing data, there is another issue about the model input. There could be a selection bias if the modelling is based solely on the accepted population with known performance. For clients that were declined in the past, the bank cannot possibly know what would have happened if they would have been accepted. In other words, the data that the bank has refers only to customer that were initially accepted for a loan. This means, that the data is already biased towards a lower default-rate. This implies that the model is not truly representative for a through-the-door client. This problem is often termed “reject inference” (Deloitte, 2016a). Reject inference is a form of missing values mechanism where the missing performance of rejected customer are "missing not at random" (MNAR) given the prior knowledge that missing is because the potential high probability of default (bad performance). The selection bias leads to a difference between accepted and rejected populations in PDs.

In order to address the selection bias, the credit risk should include both populations (accepted and rejected customers). This means that unknown performance of the rejects needs to be inferred, which is completed using the Reject inference (RI) method. There are a few extra steps required during the credit risk modelling if considering RI:

1. Build a logistic regression model on the accepts (this step has been discussed in the previous chapters)
2. Infer the rejects using a reject inference technique
3. Combine the accepts and the inferred rejects into a single dataset as a complete population
4. Build a new logistic regression model on complete population

There are two broad approaches used to infer the missing performance: assignment and augmentation, each having a different set of techniques. The most

popular techniques within the two approaches are proportional assignment, simple and fuzzy augmentation and parcelling[34].

6.2.3 Credit Risk Modelling

Generally, statistical models can provide a better understanding of the explanatory variables by estimating their parameters. Artificial Intelligent models have been popular recently and have advantages in machine learning skills which can optimise other goals and may achieve a better prediction accuracy. In light of this, it is reasonable to apply some artificial intelligence models for further studies, such as support vector machine (H. S. Kim and Sohn, 2010; S. Y. Kim, 2011), random forest classifier (Fantazzini and Figini, 2008).

6.2.4 Micro-Enterprises

According to the report from (Chakrabarty, 2010), Micro and Small-sized enterprises (MSEs) differ from the medium ones in some aspects. In the process of globalisation featured by competition and innovation, MSEs are handicapped in achieving economies of scale. MSEs more heavily rely on bank financing for purchasing land and equipment so availability of appropriate credit at reasonable rates becomes critical. Micro enterprises, from their definitions we can see, are heavily influenced by their owners since all decisions are made by less than 10 people. It is, therefore, the characteristics of the board members that are of great importance in determining the future of their business. It is similar to the argument from (Berger and Frame, 2007) and (Berger and Black, 2011) that the correlation of personal and business success need to be considered within the scope of consumer credit scoring. It is believed that credit of micro-enterprises is strongly linked to their owners' behaviour.

Future research will bridge this gap by focusing on MSEs, the subset of SMEs and build separate models to predict their credit risk. More developed models will be

[34] Source: https://www.worldprogramming.com/blog/credit_scoring_pt6

considered, and the time effect will be added to the model to make prediction dynamically. The research will be useful for creditors, such as banks in their risk assessment.

Appendix

Appendix A:

Variables selected by (M. Ma, 2016)

	Variables	Definitions
Start-ups	ref01_01	Legal form
	ref01_04	Company is subsidiary
	ref01_28	1992 SIC code
	ref01_33	Region
	ref03_03	Proportion of current directors to previous directors in the last year
	ref03_08	Oldest age of current directors/proprietors supplied (years)
	ref03_09	Number of directors holding shares
	ref05_04	Total value of judgements in the last 12 months
	ref06_03	Number of previous searches (last 12m)
	ref07_38	Time since last derogatory data item (months)
	ref10_01	Lateness of accounts
	ref10_13	Time since last annual return
	ref10_19	Total assets
Non-Start-ups	ref01_01	Legal form
	ref01_06	Parent company – derog details
	ref01_28	1992 SIC code
	ref01_33	Region
	ref03_01	No. Of ‘current’ directors
	ref03_03	Proportion of current directors to previous directors in the last year
	ref04_05	Pp worst (company DBT - industry DBT) in the last 12 months
	ref05_04	Total value of judgements in the last 12 months
	ref06_03	Number of previous searches (last 12m)
	ref07_38	Time since last derogatory data item (months)
	ref10_01	Lateness of accounts
	ref10_13	Time since last annual return
	ref10_30	Total fixed assets as a percentage of total assets
	ref10_48	Debt gearing (%)
	ref11_01	Percentage change in shareholders’ funds
	ref11_04	Percentage change in total assets

Appendix B:

Logistic Regression Results coefficient (standard error) of 50th dataset in 2009

Variables	Start-up	Non-start-up
ref01_011	-2.676*** (0.398)	-4.048*** (0.735)
ref01_012	-2.436*** (0.128)	-5.956*** (0.332)
ref01_013	-1.063** (0.537)	-6.171*** (0.579)
ref01_015	-0.667*** (0.230)	-5.639*** (0.393)
ref01_016	11.711 (92.530)	19.121 (83.359)
ref01_017	-1.431*** (0.177)	-5.981*** (0.371)
ref01_018	13.720 (486.368)	9.688*** (1.127)
ref01_042	1.718*** (0.104)	
ref01_043	2.653*** (0.478)	
ref01_019		-7.399*** (1.462)
ref01_062		-0.370*** (0.069)
ref01_063		-0.196 (0.274)
ref01_064		-0.143 (0.278)
ref01_282	0.306 (0.645)	2.078*** (0.747)
ref01_283	-0.826** (0.381)	0.756* (0.452)
ref01_284	0.186 (0.213)	0.888*** (0.221)
ref01_285	1.159** (0.465)	1.271* (0.651)
ref01_286	0.176 (0.201)	1.112*** (0.212)
ref01_287	-0.042 (0.202)	1.138*** (0.211)
ref01_288	0.112 (0.213)	0.760*** (0.223)
ref01_289	-0.126 (0.210)	0.729*** (0.219)
ref01_2810	0.193 (0.229)	1.065*** (0.238)
ref01_2811	0.117 (0.197)	1.246*** (0.208)
ref01_2812	-0.032 (0.196)	0.606*** (0.207)
ref01_2813	-0.089 (0.236)	2.007*** (0.257)
ref01_2814	0.185 (0.216)	1.187*** (0.243)
ref01_2815	0.068 (0.202)	1.049*** (0.214)
ref01_332	0.219*** (0.047)	-0.508*** (0.050)
ref01_333	0.329*** (0.068)	-0.152** (0.073)
ref01_334	0.200* (0.108)	-0.282** (0.114)
ref01_335	0.148*** (0.052)	-0.443*** (0.063)
ref01_336	0.176** (0.078)	-0.348*** (0.087)
ref01_337	0.374*** (0.057)	0.124* (0.066)
ref01_338	0.356*** (0.053)	-0.184*** (0.063)
ref01_339	0.292*** (0.064)	-0.511*** (0.071)
ref01_3310	0.415*** (0.076)	0.339*** (0.083)
ref01_3311	0.227** (0.103)	0.332*** (0.109)
ref01_3312	0.046 (0.267)	-1.949*** (0.464)
ref01_3313	-0.482 (1.223)	10.740 (399.394)
ref03_01		-0.192*** (0.024)
ref03_03	-0.448*** (0.025)	0.792*** (0.029)
ref03_08	0.016*** (0.001)	
ref03_09	1.146*** (0.027)	
ref04_05		-0.005*** (0.0002)
ref05_04	-1E-10	-0.00004*** (0.00001)
ref06_03	0.043*** (0.008)	0.004 (0.006)
ref07_38	0.384*** (0.010)	0.077*** (0.002)
ref10_01	-0.184*** (0.003)	-0.033*** (0.002)
ref10_13	-0.150*** (0.004)	-0.062*** (0.002)
ref10_19	0.00000*** (0.00000)	
ref10_30		0.011*** (0.001)
ref10_48		0.00000 (0.00000)
ref11_01		0.00000 (0.00000)
ref11_04		0.00000 (0.00000)
Constant	1.382*** (0.244)	5.561*** (0.408)
Observations	43,353	49,105
Log Likelihood	-15,055.71	-13,096.85
Akaike Inf. Crit.	30,201.42	26,293.71

Note: *p**p***p<0.01

References

Aalto, J., Pirinen, P., Heikkinen, J., & Venäläinen, A. (2012). Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models. *Theoretical and Applied Climatology, 112*(1-2), pp. 99-111. doi:10.1007/s00704-012-0716-9 Retrieved from <Go to ISI>://WOS:000316574700008

Aebi, V., Sabato, G., & Schmid, M. (2012). Risk management, corporate governance, and bank performance in the financial crisis. *Journal of Banking & Finance, 36*(12), pp. 3213-3226. doi:10.1016/j.jbankfin.2011.10.020 Retrieved from <Go to ISI>://WOS:000310393900008

Akaike, H. (1987). Factor analysis and AIC *Selected Papers of Hirotugu Akaike* (pp. 371-386): Springer.

Alfaro, E., Gámez, M., & García, N. (2018). *Ensemble Classification Methods with Applications in R*: John Wiley & Sons.

Allen, L., DeLong, G., & Saunders, A. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance, 28*(4), pp. 727-752. doi:10.1016/s0378-4266(03)00197-3 Retrieved from <Go to ISI>://WOS:000220042600002

Allison, P. (2012a). Why maximum likelihood is better than multiple imputation. *Statistical Horizons*

Allison, P. (2012b). Why you probably need more imputations than you think. *Statistical Horizons. Retrevied from http://statisticalhorizons. com/more-imputations*

Allison, P. D. (1999). Logistic regression using the SAS system: theory and application. *SAS Institute Corp., USA*

Allison, P. D. (2001). *Missing data*: Sage publications.

Allison, P. D. (2016). Multiple Imputation for Missing Data. *Sociological methods & research, 28*(3), pp. 301-309. doi:10.1177/0049124100028003003 Retrieved from <Go to ISI>://WOS:000085121000003

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance, 23*(4), pp. 589-609.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance, 1*(1), pp. 29-54.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance, 18*(3), pp. 505-529. doi:10.1016/0378-4266(94)90007-8 Retrieved from <Go to ISI>://WOS:A1994NZ38000007

Altman, E. I., & Sabato, G. (2005). Effects of the new Basel capital accord on bank capital requirements for SMEs. *Journal of financial services research, 28*(1-3), pp. 15-42.

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus-a Journal of Accounting Finance and Business Studies, 43*(3), pp. 332-357.

doi:10.1111/j.1467-6281.2007.00234.x Retrieved from <Go to ISI>://WOS:000249328700008

Altman, E. I., Sabato, G., & Wilson, N. (2008). The value of non-financial information in SME risk management.

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*

Anderson, R. (2015). *Piecewise Logistic Regression: an Application in Credit Scoring.* Credit Scoring and Control Conference XIV.

Andreeva, G., Calabrese, R., & Osmetti, S. A. (2016). A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. *European Journal of Operational Research, 249*(2), pp. 506-516. doi:10.1016/j.ejor.2015.07.062 Retrieved from <Go to ISI>://WOS:000366951100012

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review, 78*(1), pp. 40-64.

Andrikopoulos, P., & Khorasgani, A. (2018). Predicting unlisted SMEs' default: Incorporating market information on accounting-based models for improved accuracy. *The British Accounting Review, 50*(5), pp. 559-573. doi:10.1016/j.bar.2018.02.003 Retrieved from <Go to ISI>://WOS:000445304700006

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: a survey and new results. *IEEE Trans Neural Netw, 12*(4), pp. 929-935. doi:10.1109/72.935101 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18249923

Austin, P. C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Stat Med, 26*(15), pp. 2937-2957. doi:10.1002/sim.2770 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17186501

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research, 20*(1), pp. 40-49.

Baltagi, B. (2008). *Econometric analysis of panel data*: John Wiley & Sons.

Balthazar, L. (2006). From Basel 1 to Basel 3 *From Basel 1 to Basel 3: The Integration of State-of-the-Art Risk Modeling in Banking Regulation* (pp. 209-213): Springer.

Baum, J. A. C., Calabrese, T., & Silverman, B. S. (2000). Don't go it alone: alliance network composition and startups' performance in Canadian biotechnology. *Strategic management journal, 21*(3), pp. 267-294. doi:10.1002/(sici)1097-0266(200003)21:3<267::Aid-smj89>3.0.Co;2-8 Retrieved from <Go to ISI>://WOS:000085900800005

BCBS. (1988). INTERNATIONAL CONVERGENCE OF CAPITAL MEASUREMENT AND CAPITAL STANDARDS. p 13. Retrieved from https://www.bis.org/publ/bcbsc111.pdf

BCBS. (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version*: Basel Committee on Banking Supervision.

BCBS. (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems*: Basel Committee on Banking Supervision.

BCBS. (2017). *Basel III: Finalising post-crisis reforms*: Basel Committee on Banking Supervision

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, pp. 71-111.

Beck, N., Katz, J. N., & Tucker, R. (1998). Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science, 42*(4), pp. 1260-1288. doi:10.2307/2991857 Retrieved from <Go to ISI>://WOS:000075891200010

Beck, T. (2013). Bank financing for SMEs–lessons from the literature. *National institute economic review, 225*(1), pp. R23-R38.

Beck, T., Demirgüç-Kunt, A., & Pería, M. S. M. (2010). Bank Financing for SMEs: Evidence Across Countries and Bank Ownership Types. *Journal of financial services research, 39*(1-2), pp. 35-54. doi:10.1007/s10693-010-0085-4 Retrieved from <Go to ISI>://WOS:000287927700003

Behr, P., & Güttler, A. (2007). Credit risk assessment and relationship lending: An empirical analysis of German small and medium-sized enterprises. *Journal of Small Business Management, 45*(2), pp. 194-213.

Bellini, T. (2019). *How to Model and Validate Expected Credit Losses for IFRS 9 and CECL: A Practical Guide with Examples Worked in R and SAS*.

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2), pp. 3302-3308. doi:10.1016/j.eswa.2008.01.005 Retrieved from <Go to ISI>://WOS:000262178100073

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting, 28*(1), pp. 171-182. doi:10.1016/j.ijforecast.2010.08.005 Retrieved from <Go to ISI>://WOS:000299451800023

Beltratti, A., & Stulz, R. M. (2012). The credit crisis around the globe: Why did some banks perform better? *Journal of Financial economics, 105*(1), pp. 1-17.

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry, 23*(2), pp. 129-143.

Berger, A. N., & Black, L. K. (2011). Bank size, lending technologies, and small business finance. *Journal of Banking & Finance, 35*(3), pp. 724-735. doi:10.1016/j.jbankfin.2010.09.004 Retrieved from <Go to ISI>://WOS:000287059200018

Berger, A. N., & Frame, W. S. (2007). Small Business Credit Scoring and Credit Availability. *Journal of Small Business Management, 45*(1), pp. 5-22. doi:10.1111/j.1540-627X.2007.00195.x Retrieved from <Go to ISI>://WOS:000242964600002

Blum, M. (1974). Failing Company Discriminant Analysis. *Journal of accounting research, 12*(1), pp. 1-25. doi:10.2307/2490525 Retrieved from <Go to ISI>://WOS:A1974AF41700001

Bodner, T. E. (2006). Missing data: prevalence and reporting practices. *Psychol Rep, 99*(3), pp. 675-680. doi:10.2466/PR0.99.3.675-680 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17305182

Bodner, T. E. (2008a). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal, 15*(4), pp. 651-675.

Bodner, T. E. (2008b). What improves with increased missing data imputations? *Structural Equation Modeling, 15*(4), pp. 651-675.

Boone, A. L., Casares Field, L., Karpoff, J. M., & Raheja, C. G. (2007). The determinants of corporate board size and composition: An empirical analysis. *Journal of Financial economics, 85*(1), pp. 66-101. doi:10.1016/j.jfineco.2006.05.004 Retrieved from <Go to ISI>://WOS:000247807700003

Borio, C., Furfine, C., & Lowe, P. (2001). Procyclicality of the financial system and financial stability: issues and policy options. *BIS papers, 1*(March), pp. 1-57.

Borio, C., & Lowe, P. (2001). To provision or not to provision. *BIS Quarterly Review, 9*(3), pp. 36-48.

Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*.

Bruderl, J., Preisendorfer, P., & Ziegler, R. (1992). Survival Chances of Newly Founded Business Organizations. *American sociological review, 57*(2), pp. 227-242. doi:10.2307/2096207 Retrieved from <Go to ISI>://WOS:A1992HR89200007

Bushman, R., & Landsman, W. R. (2010). The pros and cons of regulating corporate reporting: A critical review of the arguments. *Accounting and Business Research, 40*(3), pp. 259-273. doi:10.1080/00014788.2010.9663400 Retrieved from <Go to ISI>://WOS:000279741500009

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, pp. 1-68.

Calabrese, R., Andreeva, G., & Ansell, J. (2017). “Birds of a Feather” Fail Together: Exploring the Nature of Dependency in SME Defaults. *Risk Analysis*

Chakrabarty, K. (2010). Bank Credit to MSMEs: Present status and way forward. *RBI monthly Bulletin*

Cho, H., Chung, J. R., & Kim, Y. J. (2014). Fixed Asset Revaluation and External Financing during the Financial Crisis: Evidence from Korea.

Cohen, B. H., & Edwards, G. (2017). The new era of expected credit loss provisioning.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*: Routledge.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods, 6*(4), pp. 330-351. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11778676

Conze, A. (2015). Probabilities of default for impairment under ifrs 9.

Coughlan, A. T., & Schmidt, R. M. (1985). Executive compensation, management turnover, and firm performance. *Journal of accounting and economics, 7*(1-3), pp. 43-66. doi:10.1016/0165-4101(85)90027-8 Retrieved from <Go to ISI>://WOS:A1985AHB3100003

Cox, D. R. (1958). The Regression-Analysis of Binary Sequences. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 20*(2), pp. 215-242. Retrieved from <Go to ISI>://WOS:A1958XF55000001

Croissant, Y., & Millo, G. (2008). Panel Data Econometrics in R: The plm Package. *Journal of statistical software, 27*(2)

Crouhy, M., Galai, D., & Mark, R. (2014). The Essentials of Risk Management. Retrieved from http://www.aafm.com.br/ebooks/AAFM%20Training%20ebook%20-%20The%20Essentials%20of%20Risk%20Management%20(Crouhy,%20Galay,%20Mark,%202009).pdf

De Andres, P., & Vallelado, E. (2008). Corporate governance in banking: The role of the board of directors. *Journal of Banking & Finance, 32*(12), pp. 2570-2580.

De Laurentis, G., Maino, R., & Molteni, L. (2011). *Developing, validating and using internal ratings: methodologies and case studies*: John Wiley & Sons.

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of accounting research, Vol. 10*, pp. 167-179.

Deloitte. (2016a). Credit scoring: Case study in data analytics. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/gx-be-aers-fsi-credit-scoring.pdf

Deloitte. (2016b). *A Drain on Resources?*

The Impact of IFRS 9 on Banking

Sector Regulatory Capita. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-fs-drain-on-resources.pdf

Deloitte. (2017). *IFRS 9: ready for impact Are you ready for the shake-up?*

Deloitte. (2015). *Fifth Global IFRS Banking Survey*. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/gx-fsi-fifth-banking-ifrs-survey-full.pdf

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1-38.

Department for Business, E. I. S. (2017). BUSINESS POPULATION ESTIMATES FOR THE UK AND REGIONS 2017. Retrieved Date from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/663235/bpe_2017_statistical_release.pdf.

Dietsch, M., & Petey, J. (2002). The credit risk in SME loans portfolios: Modeling issues, pricing, and capital requirements. *Journal of Banking & Finance, 26*(2-3), pp. 303-322. doi:10.1016/s0378-4266(01)00224-2 Retrieved from <Go to ISI>://WOS:000173944800007

Dietsch, M., & Petey, J. (2004). Should SME exposures be treated as retail or corporate exposures? A comparative analysis of default probabilities and asset correlations in French and German SMEs. *Journal of Banking & Finance, 28*(4), pp. 773-788. doi:10.1016/s0378-4266(03)00199-7 Retrieved from <Go to ISI>://WOS:000220042600004

Dionne, G. (2013). Risk management: History, definition, and critique. *Risk Management and Insurance Review, 16*(2), pp. 147-166.

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *Am J Epidemiol, 156*(3), pp. 193-203. doi:10.1093/aje/kwf062 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12142253

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), p 222. doi:10.1186/2193-1801-2-222 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23853744

Doorsselaere, J. V. (2015). Wolters Kluwer: Expected Losses Accounting Under IFRS 9. Retrieved Date from http://www.risktech-forum.com/opinion/wolters-kluwer-expected-losses-accounting-under-ifrs-9.

Duan, H., Han, X., & Yang, H. (2009). An analysis of causes for SMEs financing difficulty. *International Journal of Business and Management, 4*(6), pp. 73-75.

Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis, 7*(2), pp. 1477-1493.

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The journal of finance, 32*(3), pp. 875-900.

Enders, C. K. (2010). *Applied missing data analysis*: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological methods, 16*(1), p 1.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behav Res Ther, 98*, pp. 4-18. doi:10.1016/j.brat.2016.11.008 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27890222

Ernst&Young. (2018). *Applying IFRS Impairment of financial instruments under IFRS 9*.

Fantazzini, D., & Figini, S. (2008). Random Survival Forests Models for SME Credit Risk Measurement. *Methodology and computing in applied probability, 11*(1), pp. 29-45. doi:10.1007/s11009-008-9078-2 Retrieved from <Go to ISI>://WOS:000262535800004

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. *Methodology and computing in applied probability, 11*(1), pp. 29-45.

Federico, J., Rabetino, R., & Kantis, H. (2012). Comparing young SMEs' growth determinants across regions. *Journal of Small Business and Enterprise Development, 19*(4), pp. 575-588.

Fernandes, N., & Fich, E. (2009). Does financial experience help banks during credit crises? *Unpublished working paper, Drexel University*

Figlewski, S., Frydman, H., & Liang, W. (2012). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance, 21*(1), pp. 87-105. doi:10.1016/j.iref.2011.05.004 Retrieved from <Go to ISI>://WOS:000296166500007

Financial Stability Board, F. (2009). *Addressing Procyclicality in the Financial System.* Report of the Financial Stability Forum.

Florez-Lopez, R. (2017). Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society, 61*(3), pp. 486-501. doi:10.1057/jors.2009.66 Retrieved from <Go to ISI>://WOS:000274317700014

Fondas, N., & Sassalos, S. (2000). A different voice in the boardroom: How the presence of women directors affects board influence over management. *Global focus, 12*(2), pp. 13-22.

Fosberg, R. H. (2012). Capital structure and the financial crisis. *Journal of Finance and Accountancy, 11*, p 1.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*: Springer series in statistics New York.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version, 1*(4)

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw, 33*(1), pp. 1-22. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20808728

Frykström, N., & Jieying, L. (2018). IFRS 9 – the new accounting standard for credit loss recognition. *Economic Commentaries, 3*

Gebhardt, G. (2016). Impairments of Greek Government Bonds under IAS 39 and IFRS 9: A Case Study. *Accounting in Europe, 13*(2), pp. 169-196. doi:10.1080/17449480.2016.1208833 Retrieved from <Go to ISI>://WOS:000387229200002

Gebhardt, G. u., & Novotny-Farkas, Z. (2011). Mandatory IFRS Adoption and Accounting Quality of European Banks. *Journal of business finance & accounting, 38*(3-4), pp. 289-333.

doi:10.1111/j.1468-5957.2011.02242.x Retrieved from <Go to ISI>://WOS:000290483200001

Gilbert, L. R., Menon, K., & Schwartz, K. B. (1990). Predicting bankruptcy for firms in financial distress. *Journal of business finance & accounting, 17*(1), pp. 161-171.

Golin, J., & Delhaise, P. (2013). *The bank credit analysis handbook: a guide for analysts, bankers and investors*: John Wiley & Sons.

Gordy, M. B. (2003). A risk-factor model foundation for ratings-based bank capital rules. *Journal of financial intermediation, 12*(3), pp. 199-232. doi:10.1016/s1042-9573(03)00040-8 Retrieved from <Go to ISI>://WOS:000184892000001

Gornjak, M. (2017). Comparison of IAS 39 and IFRS 9: The Analysis of Replacement. *International Journal of Management, Knowledge and Learning, 6*(1), pp. 115-130.

Gov, U. (2018). Company accounts guidance. Retrieved Date from https://www.gov.uk/government/publications/life-of-a-company-annual-requirements/life-of-a-company-part-1-accounts#small-companies.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annu Rev Psychol, 60*, pp. 549-576. doi:10.1146/annurev.psych.58.110405.085530 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18652544

Graham, J. W. (2012). Missing data theory *Missing Data* (pp. 3-46): Springer.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of psychology*, pp. 87-114.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science, 8*(3), pp. 206-213.

Green, P. J., & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*: CRC Press.

Guest, P. M. (2009). The impact of board size on firm performance: evidence from the UK. *The European Journal of Finance, 15*(4), pp. 385-404. doi:10.1080/13518470802466121 Retrieved from <Go to ISI>://WOS:000266356000002

Gujarati, D. (2014). *Econometrics by example*: Macmillan International Higher Education.

Gupta, J., Gregoriou, A., & Ebrahimi, T. (2017). Empirical comparison of hazard models in predicting SMEs failure. *Quantitative Finance, 18*(3), pp. 437-466. doi:10.1080/14697688.2017.1307514 Retrieved from <Go to ISI>://WOS:000424952100007

Gupta, J., Wilson, N., Gregoriou, A., & Healy, J. (2014). The effect of internationalisation on modelling credit risk for SMEs: Evidence from UK market. *Journal of International Financial Markets, Institutions and Money, 31*, pp. 397-413. doi:10.1016/j.intfin.2014.05.001 Retrieved from <Go to ISI>://WOS:000338738200021

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society Series a-Statistics in Society, 160*(3), pp. 523-541. Retrieved from <Go to ISI>://WOS:A1997YE17400030

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science, 1*(3), pp. 297-318.

Hastie, T., & Tibshirani, R. (1987). Generalized Additive Models: Some Applications. *Journal of the American statistical Association, 82*(398), pp. 371-386. doi:10.1080/01621459.1987.10478440 Retrieved from <Go to ISI>://WOS:A1987J105700001

Hastie, T. J. (2017). Generalized additive models *Statistical models in S* (pp. 249-307): Routledge.

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P., & Catalano, P. (2010). Multiple imputation in a large-scale complex survey: a practical guide. *Stat Methods Med Res, 19*(6), pp. 653-670. doi:10.1177/0962280208101273 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19654173

Heitjan, D. F., & Little, R. J. A. (1991). Multiple Imputation for the Fatal Accident Reporting System. *Journal of the Royal Statistical Society Series C-Applied Statistics, 40*(1), pp. 13-29. Retrieved from <Go to ISI>://WOS:A1991EV15100002

Hodson, D., & Mabbett, D. (2009). UK economic policy and the global financial crisis: paradigm lost? *JCMS: journal of common market studies, 47*(5), pp. 1041-1061.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*(1), pp. 55-&. Retrieved from <Go to ISI>://WOS:A1970F898700005

Horowitz, J. L., & Savin, N. E. (2001). Binary Response Models: Logits, Probits and Semiparametrics. *Journal of Economic Perspectives, 15*(4), pp. 43-56. doi:10.1257/jep.15.4.43 Retrieved from <Go to ISI>://WOS:000172742200004

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat, 61*(1), pp. 79-90. doi:10.1198/000313007X172556 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17401454

Hsiao, C. (2014). *Analysis of panel data*: Cambridge university press.

Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications, 33*(2), pp. 434-440. doi:10.1016/j.eswa.2006.05.006 Retrieved from <Go to ISI>://WOS:000244344000019

Hudson, J. (1987). The age, regional, and industrial structure of company liquidations. *Journal of business finance & accounting, 14*(2), pp. 199-213.

Huian, M. C. (2012). Accounting for financial assets and financial liabilities according to IFRS 9. *Annals of the Alexandru Ioan Cuza University-Economics, 59*(1), pp. 27-47.

Hull, J. (2012). *Risk management and financial institutions,+ Web Site*: John Wiley & Sons.

Hussain, J., Salia, S., & Karim, A. (2018). Is knowledge that powerful? Financial literacy and access to finance: An analysis of enterprises in the UK. *Journal of Small Business and Enterprise Development, 25*(6), pp. 985-1003.

IASB. (2003). IAS 39: Financial instruments: Recognition and measurement.

IASB. (2014). IFRS 9: Financial instruments.

Inam, F., Inam, A., Mian, M. A., Sheikh, A. A., & Awan, H. M. (2018). Forecasting Bankruptcy for organizational sustainability in Pakistan: Using artificial neural networks, logit regression, and discriminant analysis. *Journal of Economic and Administrative Sciences*

Iqbal, A., & Kume, O. (2014). Impact of financial crisis on firms' capital structure in UK, France, and Germany. *Multinational Finance Journal, 18*(3/4), pp. 249-280.

Irwin, D., & Scott, J. M. (2010). Barriers faced by SMEs in raising bank finance. *International journal of entrepreneurial behavior & research, 16*(3), pp. 245-259.

Ivashina, V., & Scharfstein, D. (2010). Bank lending during the financial crisis of 2008. *Journal of Financial economics, 97*(3), pp. 319-338. doi:10.1016/j.jfineco.2009.12.001 Retrieved from <Go to ISI>://WOS:000279188600003

Jacobson, T., Lindé, J., & Roszbach, K. (2005). Credit risk versus capital requirements under Basel II: are SME loans and retail credit really different? *Journal of financial services research, 28*(1-3), pp. 43-75.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*: Springer.

Jarrow, R. A., & Turnbull, S. M. (1992). *The pricing and hedging of options on financial securities subject to credit risk: The discrete time case*: Queen's University, School of Business, Research Program.

Jelicic, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Dev Psychol, 45*(4), pp. 1195-1199. doi:10.1037/a0015665 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19586189

Kennedy, P. (2003). *A guide to econometrics*: MIT press.

Kesner, I. F. (1988). Directors' Characteristics and Committee Membership: An Investigation of Type, Occupation, Tenure, and Gender. *Academy of Management journal, 31*(1), pp. 66-84. doi:10.2307/256498 Retrieved from <Go to ISI>://WOS:A1988M330900003

Khorasgani, A., & Gupta, J. (2017). Estimating Reserve Requirement for Credit Portfolio of UK SMEs.

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research, 201*(3), pp. 838-846. doi:10.1016/j.ejor.2009.03.036 Retrieved from <Go to ISI>://WOS:000271261200019

Kim, S. Y. (2011). Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. *The Service Industries*

Journal, 31(3), pp. 441-468. doi:10.1080/02642060802712848 Retrieved from <Go to ISI>://WOS:000287417300009

Kirkpatrick, G. (2009). The corporate governance lessons from the financial crisis. *OECD Journal: Financial Market Trends, 1*(1), pp. 61-87.

Kohler, U., & Kreuter, F. (2005). *Data analysis using Stata*: Stata press.

Krüger, S., Rösch, D., & Scheule, H. (2018). The impact of loan loss provisioning on bank capital requirements. *Journal of Financial Stability, 36*, pp. 114-129. doi:10.1016/j.jfs.2018.02.009 Retrieved from <Go to ISI>://WOS:000434490200009

Kumar, V., Sabri, S., Garza-Reyes, J. A., Nadeem, S. P., Kumari, A., & Akkaranggoon, S. (2018). *The challenges of GSCM implementation in the UK manufacturing SMEs.* 2018 International Conference on Production and Operations Management Society (POMS).

Larsem, K. (2015). Data Exploration with Weight of Evidence and Information Value in R. Retrieved Date from https://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/.

Lazure, A. (2017). *Improving Credit Classification Using Machine Learning Techniques* (

Leathwick, J. R., Elith, J., & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling, 199*(2), pp. 188-196. doi:10.1016/j.ecolmodel.2006.05.022 Retrieved from <Go to ISI>://WOS:000241994800007

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol, 171*(5), pp. 624-632. doi:10.1093/aje/kwp425 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20106935

Lee, K. J., & Carlin, J. B. (2017). Multiple imputation in the presence of non-normal data. *Statistics in medicine, 36*(4), pp. 606-617.

Lee, N., Sameen, H., & Cowling, M. (2015). Access to finance for innovative SMEs since the financial crisis. *Research policy, 44*(2), pp. 370-380. doi:10.1016/j.respol.2014.09.008 Retrieved from <Go to ISI>://WOS:000348964800006

Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), pp. 245-254. doi:10.1016/s0957-4174(02)00044-1 Retrieved from <Go to ISI>://WOS:000178264000007

Li, K., Niskanen, J., Kolehmainen, M., & Niskanen, M. (2016). Financial innovation: Credit default hybrid model for SME lending. *Expert Systems with Applications, 61*, pp. 343-355. doi:10.1016/j.eswa.2016.05.029 Retrieved from <Go to ISI>://WOS:000379634700028

Lim, C. Y., & Yong, K. O. (2017). Regulatory pressure and income smoothing by banks in response to anticipated changes to the Basel II Accord. *China Journal of Accounting Research, 10*(1), pp. 9-32. doi:10.1016/j.cjar.2016.08.003 Retrieved from <Go to ISI>://WOS:000399338000002

Lin, S.-M. (2007a). SMEs credit risk modelling for internal rating based approach in banking implementation of Basel II requirement.

Lin, S.-M. (2007b). *SMEs Credit Risk Modelling for Internal Rating Based Approach in Banking Implementation of Basel II Requirement* (Doctor of Philosophy The University of Edinburgh

Lin, W. L., Yip, K., Sambasivan, M., & Ho, J. A. (2018). Corporate Debt Policy of Malaysian SMEs: Empirical Evidence from Firm Dynamic Panel Data. *International Journal of Economics and Management, 12*(2), pp. 491-508.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), pp. 287-296.

Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American statistical Association, 87*(420), pp. 1227-1237.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*: John Wiley & Sons.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*: John Wiley & Sons.

Lo, A. W. (1986). Logit versus discriminant analysis. *Journal of econometrics, 31*(2), pp. 151-178. doi:10.1016/0304-4076(86)90046-1 Retrieved from <Go to ISI>://WOS:A1986C545100002

Long, Q., Zhang, X., & Hsu, C. H. (2011). Nonparametric multiple imputation for receiver operating characteristics analysis when some biomarker values are missing at random. *Stat Med, 30*(26), pp. 3149-3161. doi:10.1002/sim.4338 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22025311

Longford, N. T. (2006). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician*: Springer Science & Business Media.

Longstaff, F. A., & Schwartz, E. S. (1995). A simple approach to valuing risky fixed and floating rate debt. *The journal of finance, 50*(3), pp. 789-819.

Lou, Y., Caruana, R., & Gehrke, J. (2012). *Intelligible models for classification and regression.* Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

Ma, M. (2016). *Through the crisis UK SMEs performance during the 'credit crunch'* (Doctoral Dissertation). University of Edinburgh, UK.

Ma, Y., & Lin, S. (2010). 'Credit crunch'and Small-and Medium-sized Enterprises: Aspects affecting survival. *Journal of Financial Services Marketing, 14*(4), pp. 290-300.

Marra, G., & Radice, R. (2010). Penalised regression splines: theory and application to medical research. *Stat Methods Med Res, 19*(2), pp. 107-125. doi:10.1177/0962280208096688 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18815162

Mayer, M., Resch, F., & Sauer, S. (2017). Validating Point-in-Time vs. Through-the-Cycle Credit Rating Systems.

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(1), pp. 53-71.

doi:10.1111/j.1467-9868.2007.00627.x Retrieved from <Go to ISI>://WOS:000252122500004

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The journal of finance, 29*(2), pp. 449-470.

Michaelas, N., Chittenden, F., & Poutziouris, P. (1999). Financial policy and capital structure choice in UK SMEs: Empirical evidence from company panel data. *Small Business Economics, 12*(2), pp. 113-130. doi:10.1023/a:1008010724051 Retrieved from <Go to ISI>://WOS:000079740700002

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K.-R. (1999). *Fisher discriminant analysis with kernels.* Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468).

Milliken, F. J., & Martins, L. L. (1996). Searching for Common Threads: Understanding the Multiple Effects of Diversity in Organizational Groups. *The Academy of Management Review, 21*(2), pp. 402-433. doi:10.2307/258667 Retrieved from <Go to ISI>://WOS:A1996UE77300009

Modina, M., & Pietrovito, F. (2014). A default prediction model for Italian SMEs: the relevance of the capital structure. *Applied Financial Economics, 24*(23), pp. 1537-1554.

Moons, K. G., Donders, R. A., Stijnen, T., & Harrell, F. E., Jr. (2006). Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol, 59*(10), pp. 1092-1101. doi:10.1016/j.jclinepi.2006.01.009 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16980150

Morgan, J. (1997). Creditmetrics-technical document. *JP Morgan, New York*

N. Berger, A., & F. Udell, G. (1998). The economics of small business finance: The roles of private equity and debt markets in the financial growth cycle. *Journal of Banking & Finance, 22*(6-8), pp. 613-673. doi:10.1016/s0378-4266(98)00038-7 Retrieved from <Go to ISI>://WOS:000076015500001

Nehrebecka, N. (2018). Predicting the Default Risk of Companies. Comparison of Credit Scoring Models: Logit Vs Support Vector Machines. *Econometrics, 22*(2), pp. 54-73.

Nehrebecka, N., & Polski, N. B. (2016). Approach to the assessment of credit risk for non-financial corporations. Evidence from Poland. *IFC Bulletins chapters, 41*

Novotny-Farkas, Z. (2016). The Interaction of the IFRS 9 Expected Loss Approach with Supervisory Rules and Implications for Financial Stability. *Accounting in Europe, 13*(2), pp. 197-227. doi:10.1080/17449480.2016.1210180 Retrieved from <Go to ISI>://WOS:000387229200003

Nunes, P. M., & Serrasqueiro, Z. (2012). Are young SMEs' survival determinants different? Empirical evidence using panel data. *Applied Economics Letters, 19*(9), pp. 849-855.

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of accounting research, 18*(1), pp. 109-131. doi:10.2307/2490395 Retrieved from <Go to ISI>://WOS:A1980KA68200007

Ong, C., Huang, J., & Tzeng, G. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), pp. 41-47. doi:10.1016/j.eswa.2005.01.003 Retrieved from <Go to ISI>://WOS:000228843300004

Orton, P., Ansell, J., & Andreeva, G. (2017). Exploring the performance of small- and medium-sized enterprises through the credit crunch. *Journal of the Operational Research Society, 66*(4), pp. 657-663. doi:10.1057/jors.2014.34 Retrieved from <Go to ISI>://WOS:000351561600012

Ozdemir, B. (2018). Evolution of risk management from risk compliance to strategic risk management: From Basel I to Basel II, III and IFRS 9. *Journal of Risk Management in Financial Institutions, 11*(1), pp. 76-85.

Ozili, P. K., & Outa, E. (2017). Bank loan loss provisions research: A review. *Borsa Istanbul Review, 17*(3), pp. 144-163. doi:10.1016/j.bir.2017.05.001 Retrieved from <Go to ISI>://WOS:000425013800002

Pagano, M., & Pica, G. (2012). Finance and employment*. *Economic Policy, 27*(69), pp. 5-55. doi:10.1111/j.1468-0327.2011.00276.x Retrieved from <Go to ISI>://WOS:000299099100002

Pan, Q., Wei, R., Shimizu, I., & Jamoom, E. (2014). Determining Sufficient Number of Imputations Using Variance of Imputation Variances: Data from 2012 NAMCS Physician Workflow Mail Survey. *Applied mathematics, 5*, p 3421.

Payne, G. T., Benson, G. S., & Finegold, D. L. (2009). Corporate Board Attributes, Team Effectiveness and Financial Performance. *Journal of Management Studies, 46*(4), pp. 704-731. doi:10.1111/j.1467-6486.2008.00819.x Retrieved from <Go to ISI>://WOS:000265509800006

Pearce, J. A., & Zahra, S. A. (1991). The relative power of ceos and boards of directors: Associations with corporate performance. *Strategic management journal, 12*(2), pp. 135-153. doi:10.1002/smj.4250120205 Retrieved from <Go to ISI>://WOS:A1991FA09700004

Pereira, J. M., Basto, M., & Silva, A. F. d. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance, 39*, pp. 634-641. doi:10.1016/s2212-5671(16)30310-0 Retrieved from <Go to ISI>://WOS:000387543400085

Peugh, J. L., & Enders, C. K. (2016). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of educational research, 74*(4), pp. 525-556. doi:10.3102/00346543074004525 Retrieved from <Go to ISI>://WOS:000226367900003

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation, 7*(4), pp. 353-383.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology, 27*(1), pp. 85-96.

Raj, A. P. (2016). IFRS 9 may make loans more expensive for borrowers. Retrieved Date from https://www.pwc.com/my/en/assets/press/161010-theedge-ifrs9-may-make-loans-more-expensive-for-borrowers.pdf.

Ramsay, T. O., Burnett, R. T., & Krewski, D. (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology, 14*(1), pp. 18-23. doi:10.1097/00001648-200301000-00009 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12500041

Rikkers, F., & Thibeault, A. E. (2009). A structural form default prediction model for SMEs, evidence from the Dutch market. *Multinational Finance Journal, 13*(3/4), pp. 229-264.

Rodwell, L., Lee, K. J., Romaniuk, H., & Carlin, J. B. (2014). Comparison of methods for imputing limited-range variables: a simulation study. *BMC Med Res Methodol, 14*(1), p 57. doi:10.1186/1471-2288-14-57 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24766825

Rostamkalaei, A., & Freel, M. (2015). The cost of growth: small firms and the pricing of bank loans. *Small Business Economics, 46*(2), pp. 255-272. doi:10.1007/s11187-015-9681-x Retrieved from <Go to ISI>://WOS:000368738300005

Roy, D. G., Bindya, K., & Swati, K. (2013). Basel I to Basel II to Basel III: A risk management journey of Indian banks. *AIMA Journal of Management Research, 7*(2/4), pp. 0974-0497.

Royston, P., & White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of statistical software, 45*(4), pp. 1-20. Retrieved from <Go to ISI>://WOS:000298032600001

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), pp. 581-592.

Rubin, D. B. (1987a). *Multiple imputation for nonresponse in surveys*: John Wiley & Sons.

Rubin, D. B. (1987b). Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association, 91*(434), pp. 473-489. Retrieved from <Go to ISI>://WOS:A1996UP55200008

Salfrán Vaquero, D. (2018). Multiple Imputation for Complex Data Sets.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*: CRC press.

Schafer, J. L. (1999). Multiple imputation: a primer. *Stat Methods Med Res, 8*(1), pp. 3-15. doi:10.1177/096228029900800102 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10347857

Schafer, J. L. (2003). Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica, 57*(1), pp. 19-35. doi:10.1111/1467-9574.00218 Retrieved from <Go to ISI>://WOS:000183544000003

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods, 7*(2), pp. 147-177. doi:10.1037/1082-989x.7.2.147 Retrieved from <Go to ISI>://WOS:000176079500001

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behav Res, 33*(4), pp. 545-571. doi:10.1207/s15327906mbr3304_5 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26753828

Schenker, N., & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis, 22*(4), pp. 425-446. doi:10.1016/0167-9473(95)00057-7 Retrieved from <Go to ISI>://WOS:A1996VB32000007

Schindele, A., & Szczesny, A. (2016). The impact of Basel II on the debt costs of German SMEs. *Journal of Business Economics, 86*(3), pp. 197-227.

Shukeri, S. N., Shin, O. W., & Shaari, M. S. (2012). Does board of director's characteristics affect firm performance? Evidence from Malaysian public listed companies. *International Business Research, 5*(9), p 120.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*: John Wiley & Sons.

Sieczka, P., Sornette, D., & Holyst, J. A. (2011). The Lehman Brothers effect and bankruptcy cascades. *The European Physical Journal B, 82*(3-4), pp. 257-269. doi:10.1140/epjb/e2011-10757-2 Retrieved from <Go to ISI>://WOS:000295334300007

Singh, V., Terjesen, S., & Vinnicombe, S. (2008). Newly appointed directors in the boardroom:: How do women and men differ? *European management journal, 26*(1), pp. 48-58.

Sohn, S. Y., Kim, D. H., & Yoon, J. H. (2016). Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing, 43*, pp. 150-158. doi:10.1016/j.asoc.2016.02.025 Retrieved from <Go to ISI>://WOS:000375042300013

Sohn, S. Y., Kim, H. S., & Moon, T. H. (2007). Predicting the financial performance index of technology fund for SME using structural equation model. *Expert Systems with Applications, 32*(3), pp. 890-898. doi:10.1016/j.eswa.2006.01.036 Retrieved from <Go to ISI>://WOS:000243799100016

Solanki, H. U., Bhatpuria, D., & Chauhan, P. (2016). Applications of generalized additive model (GAM) to satellite-derived variables and fishery data for prediction of fishery resources distributions in the Arabian Sea. *Geocarto international, 32*(1), pp. 30-43. doi:10.1080/10106049.2015.1120357 Retrieved from <Go to ISI>://WOS:000389047400003

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am J Epidemiol, 169*(9), pp. 1133-1139. doi:10.1093/aje/kwp026 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19318618

Suisse, C. (1997). Credit Risk: A Credit Risk Management Framework, Credit Suisse Financial Products. *New York. NY*

Summit, L. (2009). Leaders' Statement. *Guardian co. uk Thursday, 2*

Taffler, R. J. (1982). Forecasting Company Failure in the UK Using Discriminant Analysis and Financial Ratio Data. *Journal of the Royal Statistical Society. Series A (General), 145*(3), pp. 342-358. doi:10.2307/2981867 Retrieved from <Go to ISI>://WOS:A1982PL00700003

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). *Distill-and-compare: Auditing black-box models using transparent model distillation.* Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.

Temim, J. (2016). The IFRS 9 Impairment Model and its Interaction with the Basel Framework. Retrieved Date from https://www.moodysanalytics.com/risk-perspectives-magazine/convergence-risk-finance-accounting-cecl/spotlight-cecl/ifrs-9-impairment-model-interaction-with-the-basel-framework.

Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios*: OUP Oxford.

Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B. B., Rashidi, P., Pardalos, P., Momcilovic, P., & Bihorac, A. (2016). Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one, 11*(5)

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological, 58*(1), pp. 267-288. Retrieved from <Go to ISI>://WOS:A1996TU31400017

Tominac, S. B., & Vašiček, V. (2018). *The impact of IFRS 9 on loan impairments in Croatian banks.* 34th International Scientific Conference on Economic and Social Development-XVIII International Social Congress (ISC-2018).

Treiman, D. J. (2014). *Quantitative data analysis: Doing social research to test ideas*: John Wiley & Sons.

Trevor, H., Robert, T., & JH, F. (2009). The elements of statistical learning: data mining, inference, and prediction: New York, NY: Springer.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res, 16*(3), pp. 219-242. doi:10.1177/0962280206074463 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17621469

Van Buuren, S. (2012). *Flexible imputation of missing data*: CRC press.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation, 76*(12), pp. 1049-1064.

Vaněk, T., & Hampel, D. (2017). The Probability of Default Under IFRS 9: Multi-period Estimation and Macroeconomic Forecast. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 65*(2), pp. 759-776.

Venkatesh, S., & Muthiah, K. (2012). SMEs in India: Importance and contribution.

Verbano, C., & Venturini, K. (2013). Managing risks in SMEs: A literature review and research agenda. *Journal of technology management & innovation, 8*(3), pp. 186-197.

Vink, G., Frank, L. E., Pannekoek, J., & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica, 68*(1), pp. 61-90. doi:10.1111/stan.12023 Retrieved from <Go to ISI>://WOS:000331198800004

Von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological methods & research, 42*(1), pp. 105-138.

Wahba, G. (1990). *Spline models for observational data*: Siam.

Wan, Y., Datta, S., Conklin, D. J., & Kong, M. (2015). Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *J Stat Comput Simul, 85*(9), pp. 1902-1916. doi:10.1080/00949655.2014.907801 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26412909

Weed, D. L. (2005). Weight of evidence: a review of concept and methods. *Risk Anal, 25*(6), pp. 1545-1557. doi:10.1111/j.1539-6924.2005.00699.x Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16506981

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med, 30*(4), pp. 377-399. doi:10.1002/sim.4067 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21225900

Widaman, K. F. (2006). III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development, 71*(3), pp. 42-64.

Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The Journal of Financial and Quantitative Analysis, 15*(3), pp. 757-770. doi:10.2307/2330408 Retrieved from <Go to ISI>://WOS:A1980KN78900012

Wilcox, J. W. (1971). A simple theory of financial ratios as predictors of failure. *Journal of accounting research*, pp. 389-395.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist, 54*(8), pp. 594-604. doi:10.1037/0003-066x.54.8.594 Retrieved from <Go to ISI>://WOS:000081919300009

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in medicine, 27*(17), pp. 3227-3246.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials, 1*(4), pp. 368-376. doi:10.1191/1740774504cn032oa Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16279275

Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R package version, 1*, p 29.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(1), pp. 95-114. doi:10.1111/1467-9868.00374 Retrieved from <Go to ISI>://WOS:000180996800005

Wood, S. N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American statistical Association, 99*(467), pp. 673-686. doi:10.1198/016214504000000980 Retrieved from <Go to ISI>://WOS:000223857500017

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(3), pp. 495-518. doi:10.1111/j.1467-9868.2007.00646.x Retrieved from <Go to ISI>://WOS:000254954500002

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(1), pp. 3-36. doi:10.1111/j.1467-9868.2010.00749.x Retrieved from <Go to ISI>://WOS:000285970300002

Wood, S. N., & Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling, 157*(2-3), pp. 157-177. doi:10.1016/s0304-3800(02)00193-x Retrieved from <Go to ISI>://WOS:000179241300006

Wood, S. N., Pya, N., & Säfken, B. (2017). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American statistical Association, 111*(516), pp. 1548-1563. doi:10.1080/01621459.2016.1180986 Retrieved from <Go to ISI>://WOS:000391900700023

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*: MIT press.

Wu, D. D., & Olson, D. L. (2010). Enterprise risk management: small business scorecard analysis. *Production Planning & Control, 20*(4), pp. 362-369. doi:10.1080/09537280902843706 Retrieved from <Go to ISI>://WOS:000266246600007

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika, 92*(4), pp. 937-950.

Yasser, Q. R., Mamun, A. A., & Rodrigs, M. (2017). Impact of board structure on firm performance: evidence from an emerging economy. *Journal of Asia Business Studies, 11*(2), pp. 210-228. doi:10.1108/jabs-06-2015-0067 Retrieved from <Go to ISI>://WOS:000404810900006

Yuan, Y. (2011). Multiple Imputation Using SAS Software. *Journal of statistical software, 45*(6), pp. 1-25. Retrieved from <Go to ISI>://WOS:000298032800001

Ze-jing, C. X.-h. Z., & Fu-qiang, W. (2008). An Empirical Study of the Credit Risk of Listed SMEs in China Based on the KMV Model [J]. *Application of Statistics and Management, 1*, p 027.

Zeng, G. (2014). A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences, 8*(65), pp. 3229-3242.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*(3), pp. 689-722.

Zhao, Y., & Long, Q. (2017). Variable selection in the presence of missing data: imputation-based methods. *Wiley Interdisciplinary Reviews: Computational Statistics, 9*(5)

Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of accounting research, 22*, pp. 59-82. doi:10.2307/2490859 Retrieved from <Go to ISI>://WOS:A1984AGS2800006

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), pp. 301-320.